Propuesta de enriquecimiento ontológico a partir de datos textuales para el idioma español en el dominio del conflicto armado colombiano

Manuela del Pilar Gómez Suta

Universidad Tecnológica de Pereira
Facultad de Ciencias Empresariales
Pereira, Colombia
2021

Propuesta de enriquecimiento ontológico a partir de datos textuales para el idioma español en el dominio del conflicto armado colombiano

Manuela del Pilar Gómez Suta

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en Investigación Operativa y Estadística

Directores:
Ph.D. José A. Soto Mejía
Ph.D. Julián D. Echeverry Correa
MsC. Carlos Mauricio Zuluaga Ramirez

Universidad Tecnológica de Pereira
Facultad de Ciencias Empresariales
Pereira, Colombia
2021

Contenido

1. Introducción **4**

2. Planteamiento y justificación **6**

3. Marco teórico **10**

3.1. Ontología 10

3.2. Enriquecimiento ontológico 11

3.3. Aprendizaje ontológico 12

3.4. Antecedentes 14

3.4.1. Enfoque basado en datos 14

3.4.2. Enfoque basado en lingüística 15

3.5. Evaluación ontológica 17

3.6. Dominio: conflicto armado colombiano 18

4. Objetivos **22**

5. Metodología **23**

5.1. Datos textuales 23

5.2. Construcción del vocabulario 23

5.3. Extracción de conceptos y relaciones 27

5.3.1. Extracción de estructuras ontológicas 27

5.3.2. Métricas para evaluar las estructuras extraídas 31

5.3.3. Depuración de conceptos 34

6. Resultados y discusión **36**

6.1. Datos textuales 36

6.2. Construcción del vocabulario 37

6.3. Extracción de conceptos y relaciones 38

6.4. Depuración de conceptos 45

7. Conclusiones **50**

8. Futuros trabajos **52**

9. Generación de nuevo conocimiento **59**

Bibliografía **60**

Índice de cuadros

3-1. Investigaciones de aprendizaje ontológico en español. 19

5-1. Datos para ejemplificar modificación al estudio de [Meijer et al., 2014] 26

6-1. Caracteristicas de listados de términos 37
6-2. Resultados agrupamiento semántico 40
6-3. Composición de conceptos generados con algoritmo LDA 40
6-4. Características numéricas de conceptos generados con algoritmo *Directed Louvain* . . 41
6-5. Composición conceptos *Directed Louvain* 1 42
6-6. Composición conceptos *Directed Louvain* 2 43
6-7. Composición conceptos *Directed Louvain* 3 44
6-8. Comparación entre la propuesta planteada y estudios similares. 46
6-9. Características numéricas de conceptos *Directed Louvain* 3 depurados 47
6-10.Resultados de *word intrusion* para una partición *Directed Louvain* 3 47
6-11.Resultados agrupamiento semántico para conceptos *Directed Louvain* 3 depurados . 49

8-1. Investigaciones de enriquecimiento ontológico relacionadas con *word embedding*. . . . 56

Índice de figuras

3-1. Proceso de enriquecimiento ontológico según el modelo de [Zablith et al., 2013]. . . . 12

3-2. Proceso de aprendizaje ontológico. 13

5-1. Metodología. 23

5-2. Ejemplo de red dirigida de palabras ponderadas 29

6-1. Coherencia del Tópico (*Topic coherence* - TC) con top términos entre 5 a 60. 39

6-2. Coherencia del Tópico (*Topic coherence* - TC) con top términos entre 5 a 60 en conceptos *Directed Louvain* 3 originales y depurados. 45

Capítulo 1

Introducción

La ontología es un medio para representar, compartir y reutilizar conocimiento de dominio específico [Clark et al., 2012], es decir, es un artefacto para capturar información semántica a través de conceptos y relaciones entre estos con el fin de representar estructuralmente el conocimiento. Por su funcionalidad, se ha empleado durante el razonamiento automático en la Web Semántica, procesos de extracción de datos en los sistemas de recuperación de información y tareas del procesamiento del lenguaje natural [Petasis et al., 2011].

No obstante, la construcción de una ontología implica retos en la adquisición y actualización de conocimiento que son usualmente procesos manuales propensos a errores que demandan tiempo y recursos calificados [Konys, 2019]. El enriquecimiento ontológico[1] facilita superar estos retos ya que es la tarea de extender los conceptos y relaciones, además colocarlos en la posición correcta dentro de un modelo [Petasis et al., 2011]. El refinamiento busca actualizar la ontología cuando esta no explica la información de un corpus que sea del mismo dominio [Zablith et al., 2013]; por ende, la evaluación ontológica es una tarea clave para gestionar la ontología y abordar la conversión de datos textuales a construcciones ontológicas [Ali et al., 2019].

La tarea de enriquecimiento ontológico ocurre en tres etapas [Zablith et al., 2013]. Primero, se detecta la necesidad de cambio ontológico retomando técnicas del aprendizaje ontológico. La segunda etapa evalúa la congruencia entre la ontología semilla y las entidades establecidas. La última etapa valida la concisión y consistencia de la ontología examinando las ampliaciones ontológicas establecidas previamente. El estudio aquí documentado se enfoca en la primera etapa donde el aprendizaje ontológico es esencial ya que permite producir, extender o adaptar ontologías mediante la estructuración del conocimiento presente en textos [Petasis et al., 2011].

El español es el tercer idioma más usado en la web, dado que 363,684,593 hispano hablantes del mundo son usuarios de internet[2], por lo anterior, transformar los dominios de internet al español es una necesidad innegable [Ochoa et al., 2013]. Esto ha acentuado la exigencia de aplicar el aprendizaje ontológico considerando las particularidades del idioma español [Wong et al., 2012].

El modelamiento de documentos en español se ha hecho hasta ahora adaptando técnicas diseñadas para el tratamiento de texto en inglés, el cual ha sido tradicionalmente el foco en los estudios de aprendizaje ontológico. La adaptabilidad de estas técnicas es pertinente por dos razones. La primera por las diferencias lingüísticas entre estos idiomas; en comparación con el inglés, el español es un idioma con un amplio sistema de inflexión verbal, además de marcadores para género y número. La segunda razón es cuantificar la contribución de las diferentes técnicas de aprendizaje ontológico

[1] También denominado refinamiento, evolución o cambio ontológico.

[2] Datos extraídos el 26/10/2020 de `https://www.internetworldstats.com/stats7.htm`

para el tratamiento de documentos en español.

Es un desafío adaptar técnicas diseñadas originalmente para modelar texto en inglés porque no se tiene registro de alguna fuente de conocimiento estructurado o datos de entrenamiento en español que estén disponibles y faciliten la formación de estructuras abstractas como conceptos y relaciones entre estos [Dellschaft and Staab, 2008, Farreres et al., 2010]. Por ende, la evaluación de técnicas para la extracción de estructuras ontológicas abstractas es comúnmente realizada por expertos humanos o empleando fuentes de conocimiento generales como LAR-WordNet, YAGO3, Wikicorpus en español, entre otros.

La evaluación manual implica costos altos por la necesidad de acceder a un número representativo de peritos, así disminuir el sesgo de la evaluación realizada por humanos [Dellschaft and Staab, 2008]. Además, cada variación que se realice en los algoritmos de interés conlleva los mismos costos altos que la primera realización, haciendo que la calibración de parámetros y las evaluaciones a gran escala sean inviables [Wong et al., 2012].

La dificultad de evaluar el desempeño de las técnicas de aprendizaje ontológico usando textos en español, se acentúa al analizar un dominio específico como el conflicto armado colombiano. Este dominio carece de esfuerzos computacionales para generar fuentes de conocimiento estructurado. Actualmente, el único medio que captura información semántica de este dominio es el tesauro elaborado por el Centro Nacional de Memoria Histórica (CNMH) [Espinosa, 2018] de Colombia. No obstante, este instrumento no posee un lenguaje formal y no sería apropiado utilizarlo durante la extracción o validación de estructuras ontológicas abstractas porque podrían existir diferencias entre la especificidad terminológica de este instrumento y el cuerpo de textos a analizar.

A nivel internacional, una estrategia para describir comunidades ha sido la formación y validación de estructuras ontológicas. Por ejemplo, en [van Holt et al., 2012, Pfeffer and Carley, 2012] procesaron fuentes textuales para caracterizar el conflicto étnico de Sudán y hechos de la primavera árabe. Sumado a esto, la construcción de ontologías ha permitido identificar los actores del conflicto y relaciones entre estos [Hutchins and Benham-Hutchins, 2010], es así como las estructuras construidas facilitan reconocer los recursos, sentimientos y actividades de comunidades analizadas [Carley et al., 2012]. En este sentido, existe evidencia empírica sobre la pertinencia y viabilidad de encauzar investigaciones enfocadas en la formación de estructuras ontológicas para describir elementos de comunidades víctimas del conflicto.

Así mismo, la propuesta de emplear ontologías se sustenta al reconocer que desde las investigaciones de carácter cualitativo se procesan y analizan textos de fuente abierta a través de los cuales etnógrafos, antropólogos y sociólogos comprender el entorno de interés [Carley et al., 2012]. No obstante, el actual volumen de textos sugiere la necesidad de diseñar un enfoque donde el conocimiento sea esquematizado en un formato comprensible por entidades inteligentes que puedan manipular esta información [van Holt et al., 2012].

En este trabajo se adaptan técnicas del modelamiento de textos en inglés para hacer la conversión semi automática de textos en español a estructuras ontológicas como lo son términos, conceptos y relaciones entre estos. La principal contribución es la propuesta para validar las estructuras ontológicas abstractas sin emplear evaluación manual ni utilizar bases de conocimiento generales. Por lo que la propuesta es económica en cuanto al empleo de tiempo y recursos calificados, además, esta es aplicable a datos y dominios que carecen de fuentes de conocimiento estructurado. Sumado a lo anterior, este trabajo analiza el dominio del conflicto armado colombiano con el propósito de describir información de las comunidades víctimas.

Capítulo 2

Planteamiento y justificación

Las estructuras de conocimiento codifican la semántica de un dominio para que esta información sea útil durante la recuperación de información, análisis del discurso, resumen automático de textos, etc [Asim et al., 2018]. Por ende, las especificaciones conceptuales facilitan tareas que requieren conocimiento de experto [Erekhinskaya et al., 2020], por ejemplo, los lexicones proporcionan un lenguaje legible (para máquinas y personas) que permite etiquetar cuerpos de textos (*i.e.* corpus), con el fin de formalizar las definiciones que pueden ser empleadas en la producción de preguntas y respuestas automáticas.

En este sentido, existe la motivación de representar los datos textuales y su conocimiento asociado de tal manera que pueda ser procesado automática o semi automáticamente por las computadoras [Ali et al., 2019, Meijer et al., 2014, Völker et al., 2008], con el fin de liberar al ser humano del grillete que representa la recuperación, el procesamiento y la extracción de la información válida a partir de textos [Ochoa et al., 2013].

Así, las ontologías[1] son esenciales porque esquematizan los elementos y características del fenómeno que se desea comprender, reduciendo la ambigüedad al sintetizar explícitamente los conceptos y relaciones que describen el comportamiento de dicho fenómeno [Mosharraf and Taghiyareh, 2017, Völker et al., 2008]. De este modo, "las ontologías proporcionan una semántica de conocimiento comprensible para la máquina" [Ali et al., 2019, p. 2] que simplifica tareas automáticas o semi automáticas donde se procesan textos, además, favorece el intercambio de información y reutilización del conocimiento [Reyes-Ortiz, 2019].

Dada la funcionalidad y ventajas que acarrean las ontologías, han atraído mucho interés en el ámbito académico e industrial [Hlomani and Stacey, 2014a], llevando a la proliferación de estas en dominios como la bioinformática, el turismo, la ingeniería de software, la medicina, los sistemas educativos, la química, la genética, las ciencias sociales, los sistemas judiciales, etc [Ochoa et al., 2013]. Las diferentes ontologías son utilizadas en aplicaciones para las cuales no fueron diseñadas originalmente, por ende, las ontologías pueden no representar la información que permite abordar las tareas de interés [Petasis et al., 2011]. Esto implica que una ontología diseñada para un fin y dominio específico puede ser ineficiente, si es empleada para una actividad diferente a la concebida originalmente o si no considera los nuevos conceptos y relaciones de un dominio [Gillani and Ko, 2015].

Lo anterior subraya la importancia de un enfoque de investigación que respalde el ciclo de vida completo de las ontologías [Zablith et al., 2013], es decir, más allá de los pasos concernientes a la

[1] Especificaciones formales explícitas de los términos en el dominio y las relaciones entre ellos [Gruber, 1993]. Una ontología define un vocabulario común entre aquellos que necesitan compartir información en un dominio, al incluir definiciones interpretables para las máquinas [Noy and McGuinness, 2001].

construcción de la ontología, es pertinente considerar estrategias que permitan mantener y refinar las ontologías de acuerdo con los cambios en los dominios que representan o los requisitos de las aplicaciones que admiten [Kamoun and Ben Yahia, 2012].

En este orden de ideas, la tarea de enriquecimiento ontológico[2] toma un rol esencial, ya que permite extender una ontología existente con conceptos y relaciones semánticas [Petasis et al., 2011]. La importancia de esta tarea se sustenta en dos aspectos. El primero es la premisa de que el conocimiento es dinámico y cambia constantemente [Brewster et al., 2004, Hlomani and Stacey, 2014b, Knoell et al., 2017], por ende, la ontología debe responder a la naturaleza de la información que condensa y poseer algún mecanismo para actualizarse cuando existan nuevos conceptos o relaciones que describan el comportamiento del fenómeno representado [Zablith et al., 2013]. El segundo aspecto conlleva la necesidad de reutilizar el conocimiento contenido en una ontología, pues el mayor impedimento para usar ontologías es el costo de su construcción [Alfonseca and Manandhar, 2002, Cimiano and Völker, 2005, Zavitsanos et al., 2010], en consecuencia, es deseable que las ontologías ya existentes puedan ser adaptadas a las actividades de los usuarios, evitando la implementación de esquematizaciones inapropiadas [Brewster et al., 2004, Zavitsanos et al., 2010] que entorpezcan el resultado final.

En [Zablith et al., 2013] plantean tres etapas para realizar la tarea de enriquecimiento ontológico. La primera detecta la necesidad de cambio ontológico, por esta razón se retoman técnicas del aprendizaje ontológico para convertir los datos textuales en estructuras ontológicas. La segunda etapa evalúa la congruencia entre la ontología semilla y las entidades establecidas, además, establece la estrategia con la cual se agregarán las estructuras que no están presentes en la ontología semilla. La última etapa valida la concisión y consistencia de la ontología examinando las ampliaciones ontológicas establecidas previamente.

El estudio aquí documentado se enfoca en la primera etapa donde el aprendizaje ontológico es esencial porque facilita extraer el conocimiento contenido en textos, con el objetivo de representarlo en un formato legible por la máquina [Al-Aswadi et al., 2020], en este sentido, el aprendizaje ontológico permite extender o adaptar ontologías de manera semi automática mediante la obtención y estructuración del conocimiento contenido en textos [Petasis et al., 2011].

El aprendizaje ontológico a partir de textos en español ha sido vagamente tratado en comparación con la modelación de textos del lenguaje inglés. Esto puede ocurrir por la riqueza y variedad lingüística del idioma español que complejiza el procesamiento de textos. Sin embargo, esta situación es contradictoria con el actual escenario donde el español es el tercer idioma más usado en la web, considerando que de los 516,655,099 hispano hablantes del mundo, el 70.41 % son usuarios de internet y este valor ha aumentado en los últimos veinte años en 2,650.4 %[3]; por lo anterior, la "computarización de los dominios de internet al español es una verdad incuestionable" [Ochoa et al., 2013, p. 2058].

De la dificultad asociada con el modelamiento del lenguaje en español y la simultánea necesidad de estudiarlo, se podría explicar el bajo número de investigaciones que se encuentran en torno al aprendizaje ontológico a partir de datos textuales en español. Para presentar trabajos concernientes al aprendizaje ontológico en español, este documento sigue los tres enfoques presentados en [Clark et al., 2012], donde el primero está basado en datos y técnicas estadísticas para la extracción de entidades ontológicas. El segundo enfoque explota el conocimiento recopilado en datos de capacitación y recursos estructurados, con el propósito de obtener información semántica de los conceptos y relaciones no taxonómicas, es así como este enfoque integra conocimiento de expertos humanos para obtener mayor precisión en las entidades extraídas. El tercer enfoque es un híbrido entre el primero y segundo.

[2]También denominado refinamiento, evolución o cambio ontológico.

[3]Datos extraídos el 26/10/2020 de `https://www.internetworldstats.com/stats7.htm`

El primer enfoque, en el aprendizaje ontológico partiendo de textos en español, emplea la hipótesis de la bolsa de palabras (*bag-of-words*) para ponderar los términos presentes en los documentos, de este modo, filtrar aquellos que no se consideran relevantes al comparar su presencia contra un *gold standard*. La clusterización jerárquica establece grupos de términos comprendidos como conceptos [Gutiérrez-Batista et al., 2018]; además, la regresión logística permite determinar cuáles términos son semejantes a conceptos de un ontología de referencia [Farreres et al., 2010]. Las técnicas estadísticas mencionadas no extraen simultáneamente conceptos y relaciones entre estos, por esto, las relaciones emergen del análisis de co-ocurrencia entre los términos.

La validación de los términos es una tarea sencilla y ejecutable mediante *gold standard*. Esto puede ser la razón de que algunas investigaciones [Ochoa et al., 2013, Galicia-Haro and Gelbukh, 2014] consideren que los términos extraídos son equivalentes a los conceptos, aun cuando esto es incoherente con la definición de concepto como agrupación de términos. Igualmente, la carencia de corpus anotados con conocimiento ontológico (conceptos y relaciones entre estos) y altos costos que implican la construcción de estos recursos pueden ser la causa de usar términos como conceptos. En [Farreres et al., 2010] los autores validan los conceptos y relaciones al evaluar la similitud entre las estructuras ontológicas extraídas y las presentes en una ontología general, es decir, los resultados de esta evaluación dependen del contenido que recopila la fuente de conocimiento general. Por lo anterior, la debilidad de este tipo de valoración es su dependencia a las medidas comparativas (como *precision* y *recall*).

En este sentido, el primer enfoque agrupa técnicas no costosas y aplicables a datos de diferentes idiomas y dominios. Sin embargo, estas técnicas no identifican la naturaleza de los conceptos y los tipos de relaciones, en consecuencia, se producen ontologías livianas y no interpretables por la máquina. La evaluación ontológica basada en *gold standard* es sencilla y ejecutable para valorar los términos extraídos, no obstante, los datos de entrenamiento con conocimiento ontológico son costosos de construir y no están disponibles para dominios e idiomas particulares, además, para evaluar las estructuras ontológicas extraídas se emplean métricas que no cuantifican el nivel de granularidad entre el vocabulario de los textos y las fuentes de conocimiento general.

Bajo el segundo enfoque, los conceptos se extraen de fuentes de conocimiento estructurado [Alemán et al., 2019], incluso cuando las estructuras conceptuales recuperadas pueden no presentarse en los textos analizados. Para la extracción de conceptos es usual emplear bases conceptuales y recursos multilingüísticos que indican las características sintácticas de un concepto [Ali et al., 2019]. Por otro lado, algunos estudios [Aguilar et al., 2016] construyen relaciones con patrones léxicos y recursos externos que vincula las formas lingüísticas a los esquemas cognitivos de las relaciones taxonómicas y no taxonómicas [Ochoa et al., 2013]. La evaluación ontológica surge de la valoración por humanos [Aguilar et al., 2016]. La evaluación manual implica costos altos por la necesidad de acceder a un número significativo de peritos para disminuir el sesgo del trabajo realizada por humanos [Dellschaft and Staab, 2008]; además, cada variación que se realice en los algoritmos de interés conlleva los mismos costos altos que la primera realización, haciendo que la calibración de parámetros y las experimentaciones a gran escala sean inviables [Wong et al., 2012].

Desde el segundo enfoque, las estructuras ontológicas extraídas contienen detalles sobre la naturaleza y tipos de relaciones. Esto produce ontologías formales y útiles en tareas de razonamiento automático. Sin embargo, las técnicas asociadas no son sencillas de aplicar en contextos que carecen o no tienen disponibles bases conceptuales, por ejemplo, en [Ochoa et al., 2013] los autores documentan el empleo de ADESSE, que es una base de construcciones verbales del español que ofrece información semántica, no obstante, los autores no dejan a disposición este recurso. El uso de patrones léxicos y fuentes de conocimiento estáticas restringen los hallazgos a las opciones aprendidas previamente, por ende, si no se ha considerado alguna forma particular se puede perder información presente en los textos. Por las falencias mencionadas, este enfoque hace hincapié en validar las estructuras capturadas mediante la valoración con peritos. Esto vuelve costoso el proceso e impide

la escalabilidad de las técnicas y resultados asociados [Wong et al., 2012].

El tercer enfoque es una conjunción entre las técnicas del primer y segundo. Algunas propuestas plantean la extracción de conceptos a través de fuentes de conocimiento y la generación de relaciones mediante análisis de co-ocurrencia. Otros estudios reconocen conceptos a partir de la frecuencia de los términos recuperados, además, establecen las relaciones mediante patrones léxicos. Algunos trabajos [Galicia-Haro and Gelbukh, 2014] sugieren homogeneizar la representación de los documentos al reemplazar los términos de los textos por su correspondiente etiqueta de LAR-WordNet para utilizar técnicas de agrupamiento jerárquico durante la aglomeración de las etiquetas, consecuentemente construir conceptos. En [Alemán et al., 2019] los autores proponen que los conceptos sean reconocidos por expertos, para después, determinar términos similares sintáctica y semánticamente a las estructuras conceptuales. Las falencias, en la validación de las estructuras ontológicas, de los dos primeros enfoques están en el tercero ya que la evaluación por humanos y basada en *gold standard* es recurrente.

Las falencias de los tres enfoques son mayores durante el análisis de un dominio específico como el conflicto armado colombiano. La importancia y complejidad de este dominio es ampliamente reconocida, pero carece de esfuerzos computacionales para generar fuentes de conocimiento estructurado. Actualmente, el único medio que captura información semántica de este dominio es el tesauro elaborado por el CNMH de Colombia [Espinosa, 2018]. No obstante, este instrumento no posee un lenguaje formal y emplearlo como base conceptual para la extracción o validación de estructuras ontológicas abstractas podría producir una evaluación inexacta, por la imposibilidad de hacer una comparación precisa entre las estructuras ontológicas abstractas extraídas y las presentes en la fuente referencial [Galicia-Haro and Gelbukh, 2014], dada la diferencia del nivel de granularidad entre el vocabulario del corpus y las etiquetas de la base conceptual.

Los trabajos presentados, sobre aprendizaje ontológico a partir de textos en español, permiten señalar que las técnicas del primer enfoque son las apropiadas para modelar textos de idiomas y dominios que carecen de recursos lingüísticos, como el dominio del conflicto armado colombiano, no obstante, estas técnicas producen ontologías livianas pues no emplean fuentes externas que faciliten reconocer la naturaleza de los conceptos y los tipos de relaciones. Sin embargo, los hallazgos del campo han generado una creciente conciencia de las complejidades abordadas al modelar el conocimiento en estructuras ontológicas. Esto ha planteado la cuestión sobre la factibilidad de construir automáticamente una ontología formal, o por el contrario, la necesidad de abordar objetivos más pragmáticos al enfocarse en la construcción y extensión automática de ontologías livianas [Wong et al., 2012].

El estudio aquí documentado respalda el enfoque práctico de la tarea de aprendizaje ontológico porque busca generar una propuesta de enriquecimiento ontológico, enfocada en el aprendizaje ontológico considerando minimizar el esfuerzo humano y empleando técnicas escalables a otros dominios e idiomas. Este estudio emplea técnicas del primer enfoque para extraer entidades ontológicas a partir de textos en español que abordan el conflicto armado colombiano. Además, esta investigación propone una evaluación basada en tarea con el propósito de aportar a la solución sobre las técnicas para validar estructuras ontológicas abstractas. Así, el estudio no requiere *gold standard* ni bases conceptuales generales, además, no asume que los términos son equivalentes a conceptos al valorar conceptos y relaciones. Este trabajo propone la extracción y validación de estructuras ontológicas abstractas desde textos, sin la necesidad de utilizar la evaluación por humanos ni fuentes de conocimiento general. Los experimentos presentados conllevan el tratamiento de textos en español sobre el dominio del conflicto armado colombiano.

Capítulo 3

Marco teórico

Este trabajo aborda el enriquecimiento ontológico enfocado en el aprendizaje ontológico. Por lo anterior, este capítulo expone la definición de ontología (3.1), la conceptualización teórica del enriquecimiento (3.2) y aprendizaje ontológico (3.3), especificando las técnicas de evaluación que permiten validar las estructuras ontológicas aprendidas (3.5). Así mismo, se presentan antecedentes (3.4) y se describe el dominio del conflicto armado colombiano (3.6), al ser el caso de interés de este estudio.

3.1. Ontología

Adquirir conocimiento útil, proveniente de textos en lenguaje natural, conlleva una serie de retos como la dificultad para realizar consultas semánticas, elevados costos para mantener los repositorios de información y alta dependencia del factor humano al depurar la documentación buscada [Cimiano, 2006]. Así ha surgido la necesidad de construir sistemas de organización y representaciones del conocimiento que faciliten la recuperación, análisis y comprensión de información contenida en recursos textuales [Gómez-Pérez, 2004].

En este orden de ideas, ha surgido la noción de representación del conocimiento que es el estudio de cómo esquematizar información con el fin de que las entidades inteligentes (como el computador) puedan razonar a partir de dicha esquematización [Guarino et al., 2009]. Una ontología es una representación que proporciona conocimiento sofisticado para el procesamiento de tareas [Cimiano and Völker, 2005], porque es una especificación explícita de una conceptualización compartida [Gruber, 1993].

Una conceptualización es una visión simplificada y abstracta del mundo que se desea representar [Guarino et al., 2009]. Dicha conceptualización es una especificación explícita al ser declarada con individualidad mediante un lenguaje que cuenta con un vocabulario y estructura de relaciones intencionales que limitan la interpretación de la conceptualización [Gómez-Pérez, 2004]. Así, una ontología puede ser una especificación formal cuando está construida a partir de un lenguaje que posee sintaxis y semántica definida [Zavitsanos et al., 2010].

En este sentido, una ontología es un recurso que representa el modelo conceptual subyacente a un dominio en términos de los conceptos relevantes y las relaciones entre estos [Cimiano, 2006], con el objetivo de definir la semántica del fenómeno modelado [Zablith et al., 2013]. Así, en [Noy and McGuinness, 2001] señalan que emplear ontologías es beneficioso porque: i) facilitan compartir el entendimiento común de la estructura de información entre personas o agentes computacionales, ii) permiten la reutilización del conocimiento con lo que se puede extender la interoperabilidad de los

componentes de un sistema, iii) hacen explícitas suposiciones del fenómeno que representan y iv) favorecen el análisis formal del conocimiento de dominio, siendo esto muy valioso cuando se adapta la información.

Dados los beneficios nombrados, las ontologías se han convertido en un mecanismo importante en la web semántica ya que permiten agregar estructura a los datos [Ochoa et al., 2013]. En este escenario, las ontologías son distribuidas y utilizadas en aplicaciones cada vez más diversas, implicando un esfuerzo de investigación enfocado en respaldar el ciclo de vida de las ontologías, en especial el mantenimiento y evolución de estas de acuerdo con los cambios en los dominios que representan o los requisitos de las aplicaciones que soportan [Zablith et al., 2013].

Dirigir esfuerzos investigativos a la evolución ontológica se justifica al comprender que la creación de ontologías es un proceso laborioso y propenso a errores ya que la información disponible de un domino es usual que se halle en forma de texto [Petasis et al., 2011], además porque se espera que las ontologías posean una cobertura significativa del dominio y fomenten la concisión del fenómeno que representan, al determinar generalizaciones significativas y coherentes [Cimiano, 2006].

3.2. Enriquecimiento ontológico

La tarea de ampliar una ontología existente mediante análisis de texto se denomina enriquecimiento ontológico, es decir, se refiere a cambiar la estructura de la ontología al extender los conceptos y relaciones, además colocarlos en la posición correcta [Petasis et al., 2011]. El refinamiento se centra en actualizar la ontología cuando esta no es suficiente para explicar la información de un corpus que sea del mismo dominio [Zablith et al., 2013], así el objetivo del enriquecimiento ontológico es acrecentar el conocimiento esquematizado en la ontología con el fin de explicar mejor la información que se recupere en el futuro [Gillani and Ko, 2015]. Por ende, el enriquecimiento ontológico es una tarea clave para gestionar la ontología y aborda la conversión de datos textuales a construcciones ontológicas [Ali et al., 2019].

Cualquier actualización en la ontología puede acarrear consecuencias, por ejemplo, agregar un objeto ontológico tiene efecto en otras estructuras a razón que se pueden modificar relaciones o la forma en que se interpreta el fenómeno representado [Gillani and Ko, 2015]. Por tal razón, la evolución ontológica posee una metodología estándar. En [Zablith et al., 2013] los autores proponen un modelo general para evolución ontológica. Este trabajo retoma la propuesta porque representa una síntesis de las formulaciones existentes en la literatura sobre la tarea de enriquecimiento ontológico. Este modelo concibe principalmente tres etapas.

1. Detección de cambio ontológico: Corresponde a identificar conceptos y relaciones en un corpus que cubra el mismo dominio de la ontología inicial [Petasis et al., 2011]. Estas estructuras ontológicas[1] sirven para tasar la congruencia de la versión actual de la ontología con el propósito de detectar la necesidad o no de cambio ontológico.
   Esta etapa aborda la conversión de datos textuales en estructuras ontológicas con estrategias del aprendizaje ontológico. Estas estructuras son evaluadas con el fin de reconocer si representan el contenido de los textos analizados.

2. Sugerencia de cambio ontológico: Inicia con la valoración de la ontología semilla mediante la evaluación ontológica. Considera seleccionar las estructuras ontológicas extraídas y no presentes en la ontología inicial con el propósito de agregarlas [Zablith et al., 2013]. Esta etapa se ejecuta con estrategias del alineamiento ontológico.

[1] Es decir, los conceptos y relaciones [Guarino et al., 2009].

3. Validación de cambio ontológico: Consiste en filtrar las transformaciones que aun cuando fueron sugeridas, no deben agregarse a la ontología para evitar incoherencias o la violación de restricciones específicas del dominio. Esta etapa utiliza la evaluación ontológica aplicando el principio de validez [Zablith et al., 2013] que consiste en: i) valorar la concisión de la ontología, al suponer que se añaden los cambios sugeridos [Kamoun and Ben Yahia, 2012], y ii) evaluar la consistencia de la ontología, después de añadir los cambios ontológicos que han superado la valoración de concisión [Hlomani and Stacey, 2014a].

El enriquecimiento ontológico aprovecha estrategias asociadas al aprendizaje, alineación y evaluación ontológica con el propósito de que la ontología refinada posea características de congruencia, concisión y consistencia [Zablith et al., 2013]. En la Figura **3-1** está la composición de las etapas asociadas al proceso de cambio ontológico considerando que en amarillo está la tarea sobre aprendizaje ontológico, en morado aparecen las tareas de evaluación ontológica y en gris se señala la tarea de alineamiento ontológico.

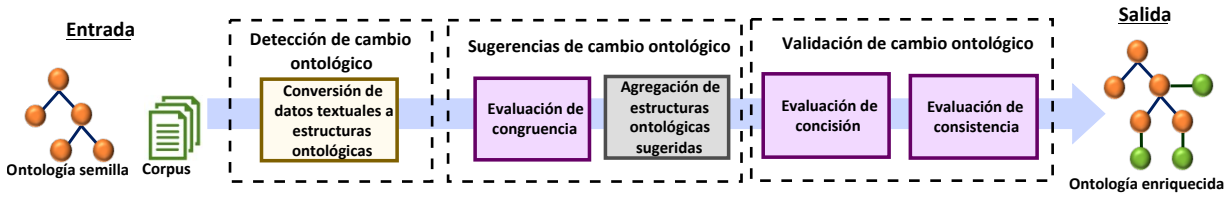

Figura **3-1**: Proceso de enriquecimiento ontológico según el modelo de [Zablith et al., 2013].

Este trabajo se enfoca en la etapa de la detección del cambio ontológico con el fin de extraer conceptos y relaciones de un corpus de texto, para luego, reconocer si estas estructuras deben ser alineadas en la ontología semilla. Por lo anterior, el siguiente apartado presenta el aprendizaje ontológico dado el innegable vínculo que tienen con el enriquecimiento ontológico y con el propósito de la investigación.

3.3. Aprendizaje ontológico

La construcción de una ontología implica retos en la adquisición y actualización de conocimiento que son procesos usualmente manuales propensos a errores que demandan tiempo y recursos calificados [Konys, 2019]. El aprendizaje ontológico sobrelleva estos retos ya que es el conjunto de métodos y técnicas para producir, extender o adaptar ontologías de manera (semi) automática mediante la obtención y estructuración del conocimiento contenido en textos disponibles en la web [Petasis et al., 2011].

El objetivo principal del aprendizaje ontológico es identificar términos, conceptos, relaciones y, opcionalmente, axiomas al analizar información textual con el propósito de construir y/o mantener una ontología [Wong et al., 2012]. La extracción de estos objetos ontológicos tiene una dependencias pues un axioma sólo surge posterior a la determinación de términos, conceptos y relaciones.

En [Buitelaar et al., 2005] los autores propusieron *Ontology Learning Layer Cake* que es la piedra angular del aprendizaje ontológico ya que describe las tareas asociadas a este proceso, en función de los elementos que conforman las estructuras de conocimiento. Los niveles o capas propuestos en [Buitelaar et al., 2005] fueron especificados por [Wong et al., 2012] en los siguientes elementos:

1. Los términos son los bloques básicos para construir cualquier estructura de conocimiento ya que son las unidades léxicas que explícitamente aparecen en el corpus, por ende, son las

representaciones textuales de los conceptos. Un término puede estar formado por un sólo token (*i.e.* victimizar) o múltiples tokens (*i.e.* justicia transicional).

2. Los conceptos son ideas abstractas que se formulan de un dominio o fenómeno [Mishra and Jain, 2015], pueden ser abstractos (*i.e.* vida libre de violencia) o concretos (*i.e.* casa). Los conceptos surgen de la agrupación de términos relacionados y, usualmente, poseen un etiqueta que representa el conglomerado. En específico, el sentido de un concepto brota del contexto formado por sus correspondientes unidades léxicas, por ejemplo, se podría usar la etiqueta *política* para las colecciones de términos *(política económica, política social, política educativa, política pública)* y *(Donald Trump, presidente, estados, gobernador, demócratas, republicanos)*, sin embargo, cada agrupación representa un concepto diferente ya que los términos del primer conjunto hacen referencia a tipos de políticas y el segundo a la política de Estados Unidos.

3. Las relaciones representan la interacción entre conceptos en una ontología [Wong et al., 2012]. Las relaciones taxonómicas facilitan construir jerarquías mediante la hiperonimia, por otro lado, las relaciones no taxonómicas describen asociaciones de dominio (*i.e.* el concepto *igualdad ante la ley* es el origen del *derecho al debido proceso*).

4. Los axiomas son preposiciones que describen un hecho siempre verdadero sobre un fenómeno permitiendo evaluar los elementos ontológicos existentes y definir restricciones para la agregación de nuevas estructuras [Petasis et al., 2011]. Los axiomas representan conocimiento empleando lenguajes formales como Web Ontology Language (OWL)[2].

La Figura **3-2** expone el proceso del aprendizaje ontológico comprendiendo la extracción de términos, formación de conceptos y relaciones, así como el establecimiento de axiomas. Este proceso describe las tareas para construir cualquier estructura de conocimiento considerando el nivel de especificidad y expresividad de cada esquematización. Por lo anterior, si el objetivo es elaborar un lexicón sólo es necesario extraer términos, por otro lado, el aprendizaje de una taxonomía implicaría llegar hasta la formación de relaciones.

Figura **3-2**: Proceso de aprendizaje ontológico.

Una ontología formal[3] hace uso intensivo de los axiomas para la especificación de información, por el contrario, una ontología liviana[4] plantea un vocabulario controlado, unidades conceptuales y relaciones taxonómicas [Wong et al., 2012]. Las ontologías formales son muy apreciadas porque facilitan tareas de razonamiento automático en la web semántica ya que brindan a los agentes computacionales la capacidad de intercambiar, reusar y compartir conocimiento sobre los conceptos y relaciones de un dominio [Guarino et al., 2009].

No obstante, el aprendizaje automático de ontologías formales es dependiente de la disponibilidad de bases de conocimiento estructurado [Clark et al., 2012]. Esta tarea es factible de realizar en dominios y lenguajes ampliamente estudiados donde ya existan estos recursos, de lo contrario, el proceso se torna semi automático o manual al emplear expertos humanos quienes señalan las propiedades del fenómeno.

[2] `https://www.w3.org/TR/owl-ref/`
[3] También llamada pesada, *heavyweight ontology* o *heavy ontology*.
[4] También llamada informal, *lightweight ontology* o *light ontology*.

Utilizar conocimiento humano implica altos costos, reduce la escalabilidad de los resultados y limita el procesamiento de datos [Dellschaft and Staab, 2008]. Por ende, la construcción automática de ontologías livianas es una alternativa económica para modelar conocimiento, además es apropiada para contextos carentes de recursos estructurados [Clark et al., 2012]. La sección 3.4 caracteriza estudios de aprendizaje ontológico desde textos en español dada la naturaleza de esta investigación.

3.4. Antecedentes

Existen diferentes encuestas que han caracterizado los sistemas de aprendizaje ontológico. En [Asim et al., 2018, Wong et al., 2012, Al-Aswadi et al., 2020] detallan técnicas para la construcción de estructuras de conocimiento, por otro lado, los autores en [Clark et al., 2012, Zablith et al., 2013] describen la relación entre el aprendizaje ontológico, la recuperación de información y el enriquecimiento ontológico. Así mismo, algunas revisiones se enfocan en dominios particulares [Petasis et al., 2011], herramientas computacionales para aprender ontologías [Mishra and Jain, 2015] ó estrategias para generar esquematizaciones específicas [Wang et al., 2017].

El Cuadro **3-1** expone algunas investigaciones de aprendizaje ontológico siguiendo la propuesta de [Clark et al., 2012]. Esta clasificación es retomada porque esquematiza de forma sencilla las técnicas para construir los elementos ontológicos, igualmente, esta propuesta ha sido utilizada por diferentes autores [Qiu et al., 2018, Meijer et al., 2014, Zouaq et al., 2011, Thenmozhi and Aravindan, 2016, Sung et al., 2008, Zafar et al., 2017], por ende, el trabajo del capítulo 5 es comparable con estudios previos. En concreto, en [Clark et al., 2012] determinan que las técnicas de aprendizaje ontológico pueden basarse en datos o explotar información lingüística, así mismo, las investigaciones normalmente emplean técnicas de uno o dos enfoques para construir las estructuras de conocimiento, por ende, tienen un planteamiento híbrido.

Adicionalmente, el Cuadro **3-1** retoma la propuesta de [Al-Aswadi et al., 2020] sobre los niveles de intervención del usuario durante la construcción de una ontología, con el propósito de cuantificar los recursos necesarios de cada investigación. Específicamente, los autores en [Al-Aswadi et al., 2020] señalan que el aprendizaje ontológico puede ser: i) manual cuando los humanos expertos establecen cada elemento de la ontología, ii) cooperativo dado que la mayoría de las tareas son realizadas por expertos o iii) semi - automático cuando la extracción de elementos ontológicos es automática y el conocimiento de peritos permite evaluar los resultados.

El Cuadro **3-1** presenta las investigaciones expuestas en las siguientes subsecciones considerando la clasificación y parámetros mencionados previamente.

3.4.1. Enfoque basado en datos

Este enfoque recopila técnicas provenientes de los sistemas de recuperación de información y minería de datos [Clark et al., 2012]. Estas técnicas consideran la información semántica del corpus y descartan conocimiento contenido en bases conceptuales de carácter general. En este orden de ideas, estas técnicas han sido ampliamente utilizadas para extraer términos, conceptos y relaciones taxonómicas, no obstante, son ineficientes durante la identificación de objetos ontológicos complejos como relaciones no taxonómicas y axiomas [Wong et al., 2012].

La ponderación estadística es una técnica que establece unidades léxicas relevantes al considerar la frecuencia con que estas aparecen dentro del corpus. Las métricas más utilizadas durante el aprendizaje ontológico son TF-IDF y TF-entropia [Alemán et al., 2019]. Así mismo, las medidas como C/NC-Value son usadas para extraer colocaciones porque tienen en cuenta la información sintáctica y la anidación de un término candidato en otros [Wong et al., 2012], por un lado, NC-Value

asigna un valor alto a las colocaciones con vecindades formadas por adjetivos, verbos y sustantivos, mientras que C-Value pondera los unigramas que componen las colocaciones al designar un peso mayor a términos frecuentes en el corpus y escasos dentro de otras palabras [Ochoa et al., 2013].

Igualmente, la técnica de análisis contrastivo evalúa la aparición de los términos en el corpus y en textos de otros dominios para identificar aquellos fenómenos de carácter general. En específico, esta técnica es una heurística basada en el hecho de que terminología genérica y dependiente del lenguaje se distribuye de forma semejante en diferentes documentos, por otro lado, características de dominio poseen comportamientos particulares [Al-Aswadi et al., 2020]. La métricas más popular del análisis contrastivo es *Domain Pertinence* [Meijer et al., 2014].

Las técnicas de ponderación estadística y análisis contrastivo son complementarias porque facilitan extraer unidades léxicas relevantes dentro del corpus y el dominio [Al-Aswadi et al., 2020], no obstante, estas técnicas producen términos dispersos difíciles de agrupar para la formación de conceptos [Wong et al., 2012]. En este orden de ideas, las técnicas para reducir la dimensionalidad son apropiadas ya que facilitan identificar sinónimos y desambiguar el sentido de las palabras [Gutiérrez-Batista et al., 2018]. El modelo de análisis latente semántico es muy utilizado porque facilita agrupar términos similares, partiendo del supuesto de que existe una estructura subyacente a la coocurrencia de unidades léxicas [Deerwester et al., 1990]. Igualmente, en [Meijer et al., 2014] establecen el sentido de los términos al calcular la similitud de estos y sus palabras contextuales, así las palabras c_j son asociadas al término t_i con el que poseen la mayor semejanza sintáctica.

Las técnicas de reducción son usualmente un paso intermedio para la formación de conceptos a través del agrupamiento que permiten establecer estructuras de conocimiento en forma de árboles, es decir, grafos sin ciclos a través de i) la detección de candidatos conceptuales, ii) la asociación de conceptos similares mediante algoritmos de agrupamiento y iii) la denominación de cada asociación reconocida [Wong et al., 2012, Gutiérrez-Batista et al., 2018].

Por otro lado, el análisis de coocurrencia identifica relaciones taxonómicas entre conceptos que tienden a ocurrir en ventanas de palabras de igual tamaño; dicha ventana puede ser un número determinado de términos, una oración, párrafo o documento [Alemán et al., 2019]. Por ejemplo, en [Meijer et al., 2014] establecen que si t_i aparece en al menos la proporción λ de todos los textos en los que aparece t_j, entonces t_i tiene un relación de subsunción con t_j. La ventaja del análisis de coocurrencia es la identificación de relaciones implícitas entre los conceptos, por ende, estas técnicas brindan un alto *recall*, no obstante, su falencia es la baja precisión ya que relaciones irrelevantes dentro del dominio son extraídas. Por lo anterior, es recomendable validar la presencia de las relaciones generadas con recursos externos.

Las técnicas del enfoque estadístico se fundamentan en que la semántica de los elementos ontológicos debe surgir de los textos sin recurrir a ninguna base de conocimiento [Zouaq et al., 2011]. Esta postura es económica además permite escalabilidad de los experimentos y resultados, no obstante, los objetos ontológicos aprendidos tienden a ser poco precisos ya que las herramientas de extracción no contienen todas las minuciosidades del dominio, por ejemplo, son incapaces de formar relaciones no taxonómicas a menos que sean conjugadas con técnicas lingüísticas. Por lo anterior, los estudios usualmente utilizan simultáneamente el enfoque basado en datos y lingüística durante el aprendizaje ontológico.

3.4.2. Enfoque basado en lingüística

Las técnicas lingüísticas explotan características del lenguaje para extraer las entidades ontológicas [Asim et al., 2018], en este sentido, identifican el rol semántico de cada unidad léxica mediante herramientas del procesamiento del lenguaje natural y recursos externos como conocimiento de

experto, lexicones, repositorios lingüísticos, bases conceptuales, etc [Al-Aswadi et al., 2020].

Las técnicas basadas en etiquetado POS emplean definiciones preestablecidas de las reglas que describen términos y conceptos potenciales de un dominio. Existen reglas generales comúnmente aceptadas, por ejemplo, que las unidades léxicas están conformadas por palabras con etiquetas *NN** [Meijer et al., 2014]. Esto permite recuperar conceptos del tipo *acuerdo paz/NN NN*, así mismo, prescinde de constructos como *seguridad ciudadana/NN ADJ*. Por lo anterior, humanos capacitados revisan las reglas generales para adaptarlas al dominio o analizan recursos externos con el fin de extraer la sintaxis de unidades conceptuales [Alemán et al., 2019]. En este sentido, las técnicas basadas en etiquetado POS producen vocabularios precisos con términos y conceptos característicos del dominio, sin embargo, el nivel de *recall* es bajo porque sólo información explícita de los textos es recuperada [Zouaq et al., 2011].

Sumado a esto, las investigaciones reportan utilizar los patrones léxicos que son las plantillas sintácticas que capturan las posibles estructuras donde aparecen los conceptos [Ochoa et al., 2013], por ejemplo, el patrón NN_1 *y* NN_2 *ADJ* permite capturar *seguridad ciudadana* de la oración *seguridad y bienestar ciudadano*.

Los recursos externos son útiles para enriquecer semánticamente la información del corpus, de esta forma, la labor manual disminuye durante la generación de reglas o patrones léxicos [Wong et al., 2012]. WordNet es el recursos externos de carácter general más popular porque organiza conceptos y términos jerárquicamente, por ende, establece posibles representaciones y características de un concepto facilitando la extracción de elementos ontológicos [Galicia-Haro and Gelbukh, 2014]. En [Gutiérrez-Batista et al., 2018] reportan homogeneizar los términos del corpus a partir de su correspondiente etiqueta de LAR - WordNet con el fin de disminuir la dimensionalidad de los objetos aprendidos. Así mismo, en [Alemán et al., 2019] utilizan recursos multilingüe para alinear el texto y recuperar entidades ontológicas.

No obstante, utilizar recursos externos como única fuente no es apropiado porque los dominios específicos poseen terminología especializada que no aparece en repositorios generales [Zouaq et al., 2011], por ejemplo, términos como *retorno de víctimas* y *restitución de derechos territoriales* no están en LAR-WordNet. Igualmente, los recursos externos pueden incluir conceptos y términos irrelevantes a una estructura de conocimiento específico [Ali et al., 2019], esto conlleva que el aprendizaje ontológico sea más complejo porque implica tareas de pos-procesamiento.

Pocas investigaciones, que examinan el idioma español, abordan la construcción de axiomas dado que esta tarea necesita expertos humanos que faciliten construir y validar los constructos generados. Por ejemplo, en [Alemán et al., 2019] utilizan expertos humanos que analizan conceptos y relaciones entre estos, después, los peritos de forma conjunta establecen los axiomas que describen estas entidades ontológicas.

Las técnicas del enfoque lingüístico permiten extraer elementos ontológicos precisos y propios de un dominio al explotar características semánticas así como el conocimiento de expertos o contenido en repositorios especializados. Más aún en [Zafar et al., 2017] los autores señalan que los sistemas ontológicos basados en lingüística superan a otros enfoques cuando los objetos ontológicos son poco frecuentes en los textos. Adicionalmente, este enfoque es el más prometedor para aprender relaciones no taxonómicas y axiomas a través de patrones léxicos y el empleo de recursos externos; sin embargo, la labor humana es incondicional para formar las plantillas de extracción y validar los resultados [Alemán et al., 2019].

Ninguna de las técnicas señaladas es útil de forma aislada, por ende, las investigaciones suelen establecer una combinación óptima de herramientas en relación a sus recursos y la tarea abordada. Una falencia del enfoque basado en lingüística es el bajo *recall* asociado a que las técnicas dependen de conocimiento previamente aprendido, es decir, información que no haga parte del conjunto

de entrenamiento no será recuperada produciendo la extracción de un número bajo de términos, conceptos y relaciones. Lo anterior puede solucionarse al conjugar técnicas lingüísticas y estadísticas como las expuestas previamente.

Los resultados del aprendizaje semi automático deben ser evaluados porque existe una brecha entre los constructos en lenguaje natural extraídos y las abstracciones conceptuales del dominio [Cimiano, 2006]. De allí la necesidad de valorar las estructuras ontológicas construidas mediante técnicas y estrategias propias de la evaluación ontológica [Zavitsanos et al., 2010]. El siguiente apartado presenta la evaluación ontológica de estructuras aprendidas de textos.

3.5. Evaluación ontológica

La evaluación ontológica es el conjunto de técnicas que verifican y validan las estructuras ontológicas aprendidas [Gómez-Pérez, 2004]. La verificación confirma la calidad de las estructuras respecto a criterios particulares y la validación comprueba que el modelo sea correcto en relación al fenómeno del mundo que representa [Petasis et al., 2011]. Existen cuatro enfoques de evaluación para cumplir con estos propósitos.

- Evaluación a partir de *gold standard*: Consiste en evaluar la congruencia entre las estructuras aprendidas contra un referencia ontológica considerada correcta. Esta referencia es usualmente construida por expertos humanos. Este enfoque utiliza las medidas de comparación (*precision* y *recall*) de los sistemas de recuperación de información [Völker et al., 2008].
- Evaluación humana: Implica la valoración de las estructuras ontológicas en relación con las destrezas y conocimiento de un experto [Petasis et al., 2011]. Los criterios de evaluación son: i) la adecuación cognitiva (*i.e.* alineamiento entre la semántica, las estructuras ontológicas y el fenómeno representado), ii) la explicabilidad, es decir, si las estructuras poseen un lenguaje que permita al experto interpretarlas y iii) la expresividad (*i.e.* el número de preguntas que los humanos pueden contestar al utilizar los datos de las estructuras ontológicas) [Degbelo, 2017].
- Evaluación basada en datos: Este enfoque parte de que la ontología es una especificación aproximada de un dominio [Guarino et al., 2009], por ende, la evaluación debe reflejar el grado de dicha aproximación al comparar las estructuras aprendidas y datos del dominio [Hlomani and Stacey, 2014a]. Estos datos deben ser representativos del dominio para medir el ajuste entre la ontología y el corpus [Petasis et al., 2011].
- Evaluación basada en tarea: Las estructuras ontológicas son utilizadas en un sistema integrado cuyo desempeño es evaluado, por ende, la valoración del modelo de conocimiento es implícita al rendimiento de la aplicación [Völker et al., 2008]. Los criterios tasados son dependientes de la tarea, por ejemplo, si la ontología es evaluada durante el razonamiento automático es adecuado cuantificar la corrección de las respuestas y el tiempo que tarda cada consulta.

El Cuadro **3-1** expone el tipo de evaluación que plantean los trabajos relacionados con el aprendizaje ontológico a partir de textos en español, de esta forma, es apropiado afirmar que la mayoría de estudios utiliza la evaluación basada en *gold standard*. Los resultados de este tipo de evaluación son discutibles porque las medidas de *precision* y *recall* son incapaces de detectar fenómenos como la sinonimia y polisemia [Sfar et al., 2016] que afectan el nivel de granularidad de las estructuras extraídas y la referencia ontológica. Por ende, esta evaluación no establece una comparación exacta entre los objetos aprendidos y el modelo estimado [Galicia-Haro and Gelbukh, 2014, Gutiérrez-Batista et al., 2018].

Igualmente, la evaluación por humanos es ampliamente aplicada pues contrastan las estructuras aprendidas y el conocimiento de expertos [Clark et al., 2012]. En [Petasis et al., 2011] señalan que estos enfoques brindan una visión concreta de la ontología aprendida. Sin embargo, esta afirmación es debatible porque los enfoques están atados a la competencia de las personas que realicen la validación [Knoell et al., 2017]. Una opción es acceder a varios peritos para reducir la subjetividad de los resultados [Dellschaft and Staab, 2008], no obstante, esto conlleva altos costos dado que el personal requerido debe ser versado en el dominio.

Las conclusiones de la evaluación basada en datos depende de la fuente de información utilizada para comparar las estructuras ontológicas, en este sentido, surgen cuestionamientos sobre la elección de los datos y cómo establecer si son representativos o no [Petasis et al., 2011]. Por consiguiente, es usual detallar la estrategia a través de la cual se recuperan los textos o mejor aún hacerlo a través de un sistema automatizado y validado. En [Völker et al., 2008] señalan dos inconvenientes de la evaluación basada en tarea. El primero establece que el rendimiento de las estructuras ontológicas puede responder al uso particular dentro de una tarea específica, por lo que no se puede generalizar el comportamiento identificado. El segundo inconveniente apunta a que la evaluación esta subordinada a información particular del dominio donde ocurre la tarea, por ejemplo, en [Liu and Alsaadi, 2020] validan el vocabulario que describe las condiciones económicas del mercado al cuantificar la tasa de rendimiento alcanzada en modelos de inversión construidos con la terminología, por ende, la evaluación está supeditada a que los datos faciliten generar modelos financieros.

Para responder al primer inconvenientes, en [Dellschaft and Staab, 2008] indican dos criterios para filtrar la intervención de la tarea y obtener conclusiones válidas sobre las técnicas comparadas. Estos criterios señalan que la evaluación debe: i) permitir la caracterización de las técnicas por lo cual es necesario emplear diferentes criterios de medición que posibiliten sopesar las fortalezas y debilidades de los procedimientos, además ii) garantizar evaluaciones a bajo costo para asegurar la valoración frecuente y a gran escala de distintos escenarios experimentales. El segundo inconveniente es evitable si los investigadores seleccionan tareas independientes de información contextual.

La caracterización de los enfoques para validar ontologías posibilita afirmar que existen diversas estrategias durante la evaluación ontológica, dado que cada investigación utiliza un enfoque dependiendo de sus recursos y el objetivo de la representación extraída [Degbelo, 2017]. Por ejemplo, en [Hicks, 2017] determinan que la evaluación varia si la ontología es vista como: i) modelo que expresa las posibles interpretaciones de un objeto, ii) sistema lógico cuya teoría axiomática describe conocimiento, ó iii) recurso comunitario y transversal a diferentes dominios y/o tareas.

La diversidad de la evaluación ontológica ha dificultado la formación y apropiación de un marco común a través del cual comparar la efectividad de las técnicas para aprender ontologías [Völker et al., 2008, Degbelo, 2017, Clark et al., 2012]. Este trabajo, consciente de este desafió y el compromiso de brindar una propuesta equiparable a estudios similares, estableció las estrategias para evaluar las estructuras ontológicas aprendidas considerando cómo investigaciones semejantes han efectuado esta tarea.

3.6. Dominio: conflicto armado colombiano

La situación de violencia en Colombia posee un eje político que se encuentra ligado a la acción de las guerrillas revolucionarias y de las fuerzas que las enfrentan [Kalyvas, 2001], sin embargo, también existen otras dimensiones que interfieren en ello como son: i) el cultivo y tráfico de droga [García, 2014], ii) el crecimiento de las bandas armadas [Alzate, 2010] y iii) la desorganización social que favorece la violencia [Blair, 2012].

Cuadro **3-1**: Investigaciones de aprendizaje ontológico en español.

Estudio	**Entrada**	**Elementos aprendidos**					**Descripción enfoque**	**Tipo de evaluación**	**Intervención usuario**
		Términos	**Conceptos**	**Relaciones taxonómicas**	**Relaciones no taxonómicas**	**Axiomas**			
[Gutiérrez-Batista et al., 2018]	Textos LAR-WordNet	×	×	×			Etiquetado Pos Recursos externos Agrupamiento	Evaluación basada en tareas	Semi automática
[Farreres et al., 2010]	Textos		×				Regresión logística Humanos expertos	Evaluación basada en humanos	Cooperativa
[Ochoa et al., 2013]	Textos ADESSE	×	×	×	×		Recursos externos Ponderación estadística Análisis contrastivo Patrones léxicos	*Gold standard*	Semi automática
[Galicia-Haro and Gelbukh, 2014]	Textos LAR-WordNet	×	×	×			Etiquetado Pos Recursos externos	*Gold standard*	Semi automática
[Meijer et al., 2014]	Textos LAR-WordNet	×	×	×			Ponderación estadística Análisis contrastivo Reducción dimensionalidad Análisis de coocurrencia	*Gold standard*	Semi automática
[Ali et al., 2019]	Textos Recurso multilingüe	×	×	×	×		Etiquetado Pos Recursos externos Análisis de coocurrencia	*Gold standard*	Semi automática
[Alemán et al., 2019]	Textos	×	×	×	×	×	Etiquetado Pos Humanos expertos Análisis de coocurrencia	Evaluación basada en humanos	Cooperativa
[Aguilar et al., 2016]	Textos	×	×	×	×		Etiquetado Pos Patrones léxicos Recursos externos Humanos expertos	Evaluación basada en humanos	Cooperativa
Propuesta	Textos	×	×	×			Etiquetado Pos Ponderación estadística Análisis contrastivo Detección de comunidades Filtrado de conceptos	Evaluación basada en tarea	Semi automática

Estos aspectos han producido dinámicas que escapan al control del Estado, debilitando sus estructuras y produciendo que su autoridad desaparezca en gran parte del territorio [Blair, 2012], ya que dichas instituciones y autoridades locales son las encargadas de edificar la agenda pública, mediante: i) la caracterización del escenario [Alzate, 2010], ii) la formación de una visión compartida y ética de gobernanza que incluya las diferentes facciones de la sociedad [Alzate and Romo, 2014], con el fin de iii) definir las estrategias y arreglos que se deben efectuar para coordinar y armonizar las políticas y recursos hacia los fines comunes del desarrollo sostenible y la paz [Blair, 2012].

Adicional a la desaparición de la autoridad del Estado, la naturaleza de las instituciones y autoridades locales está guiadas por un modelo de gestión pública, cuya premisa es satisfacer las necesidades de los ciudadanos a través de la efectividad en el diseño y la aplicación de políticas, mediante la profesionalización de la administración de lo público [Christensen and Laegreid, 2005]. En este orden de ideas, la edificación de la agenda pública queda enmarcada en las competencias de los funcionarios delegados [Christensen and Laegreid, 2005], cuando se esperaría que este sea el espacio en el cual diversos actores sociales, vinculados o no al gobierno, evidencien sus posiciones, argumentos y necesidades frente a una misma situación [Alzate and Romo, 2014].

Lo anterior obstaculiza el proceso de formación de la política pública, pues como es señalado en [Kingdon, 1984], durante éste es preponderante reconocer y analizar: i) los elementos del entorno social y político que influyen en la configuración de una política pública y ii) las preferencias, creencias e intereses de los actores involucrados. El Estado consciente de su responsabilidad ha buscado herramientas que le permitan describir la comunidad de interés a la luz de la formación de la agenda pública. En este orden de ideas, a las instituciones estatales se les han ofrecido herramientas cualitativas y estadísticas que les ayudan a obtener información sobre la comunidad de interés y sus problemáticas.

No obstante, estos insumos aún siendo valiosos para la comprensión profunda del medio sociocultural, presentan una serie de falencias: los proyectos antropológicos, sociológicos y etnográficos cualitativos requieren masivas investigaciones detalladas sobre una región o grupo, pueden llevar años y sus resultados no son escalables porque los métodos utilizados son a menudo específicos del contexto [Carley et al., 2012]; por otro lado, las investigaciones estadísticas parten del análisis de observaciones en un solo punto en el tiempo o de pocas en un intervalo limitado [Ilgen and Hulin, 2000].

Desde el marco nacional colombiano, no se identifican propuestas que posibiliten a las instituciones estatales describir el entorno de la comunidad que será afectada por la política pública, aun cuando existe un tesauro elaborado por el CNMH con el fin de facilitar la recuperación y análisis de información relacionada con las graves violaciones a los derechos humanos e infracciones al derecho humanitario con ocasión del conflicto armado colombiano [Espinosa, 2018].

En el escenario internacional investigadores, analistas y encargados de describir comunidades hacen uso de ontologías que facilitan la comprensión rápida sobre la disposición de la comunidad afectada y los recientes cambios en el entorno sociocultural [Carley et al., 2012]. Por ejemplo, en [van Holt et al., 2012] caracterizaron el conflicto étnico de Sudán entre 2003-2010 a través de una ontología construida con fuentes textuales, de esta forma, los autores comprobaron que la situación de paz para este país estaba condicionada a la relación entre la posición geoespacial de los grupos étnicos y recursos del medio ambiente. Igualmente, en [Pfeffer and Carley, 2012] procesaron información proveniente de periódicos, con el objetivo de construir una ontología con la cual describieron y comprendieron las causas de acontecimientos durante la primavera árabe.

Bajo esta misma línea, en [Carley et al., 2012] generaron una ontología que caracterizaba la dinámica de Afganistán durante el 2010 – 2012, de esta manera, los autores brindaron una herramienta con la cual se podían responder preguntas sobre quiénes eran los actores clave, cuáles eran los temas relevantes, sentimientos, recursos y actividades de la comunidad, además identificar qué

papel desempeñaban los diversos actores. De igual modo, las ontologías formadas desde texto han sido aprovechadas para reconocer la dinámica de grupos de interés, ejemplo de está en [Hutchins and Benham-Hutchins, 2010] donde realizaron un análisis multimodal de datos estructurados y no estructurados, para entender las relaciones entre las personas que conformaban la red criminal de narcotráfico en parte de los Estados Unidos de América, así como los procesos de financiamiento, suministro de transporte y reclutamiento.

En [van Holt et al., 2013] evaluaron el enfoque de ontología en comparación al análisis manual que realizaron un grupo de expertos sobre textos relacionados a la formación de la sociedad somalí. Los autores emplearon como métricas de evaluación el número de conceptos que se recuperaban, la exactitud de estos constructos y el tiempo empleado en la tarea, de esta forma, los investigadores demostraron que la codificación semi automática de textos para producir una ontología tuvo el mayor equilibrio entre los estándares valorados.

Los trabajos previamente comentados, son evidencia empírica de que la ontología es un adecuado mecanismo para describir los elementos del entorno social y actores que interactúan en la formación de políticas públicas, en este sentido, se justifican que la presente investigación haya decidido generar una propuesta de enriquecimiento ontológico enfocada en aprendizaje ontológico considerando como dominio el conflicto armado colombiano.

Capítulo 4

Objetivos

Los trabajos presentados en la sección 3.4 permiten señalar que las técnicas del enfoque basado en datos (ver sección 3.4.1) son las apropiadas para modelar textos de idiomas y dominios que carecen de recursos lingüísticos, como es el caso de este trabajo orientado a el español y dominio del conflicto armado colombiano. Además este enfoque es adecuado porque la investigación busca generar una propuesta de enriquecimiento ontológico minimizando el esfuerzo humano y empleando técnicas escalables.

Es así como este estudio propone una evaluación basada en tarea con el propósito de aportar a la solución sobre las técnicas para validar estructuras ontológicas abstractas, por lo anterior, el trabajo no requiere *gold standard* ni bases conceptuales generales para validar los conceptos construidos. Sumado a esto, la investigación no asume que los términos son equivalentes a unidades conceptuales y no valora las estructuras ontológicas mediante una comparación inexacta con las etiquetas de una ontología general. En este sentido, se propone la extracción y validación de estructuras ontológicas abstractas desde textos, sin la necesidad de utilizar la evaluación basada en humanos ni fuentes de conocimiento general. Los experimentos presentados conllevan el tratamiento de textos en español sobre el dominio del conflicto armado colombiano. En este sentido, los objetivos son:

Objetivo general

Generar una propuesta de enriquecimiento ontológico enfocada en el aprendizaje ontológico, a partir de datos textuales para el idioma español en el dominio del conflicto armado colombiano.

Objetivos específicos

1. Preparar los datos textuales que describen las graves violaciones a los derechos humanos e infracciones al derecho humanitario con ocasión del conflicto armado colombiano.
2. Formular algoritmos para el aprendizaje ontológico considerando las limitaciones de recursos lingüísticos que tiene el idioma español y el dominio del conflicto armado colombiano.
3. Evaluar las estructuras ontológicas que sean producto de aplicar los algoritmos establecidos.

Capítulo 5

Metodología

La Figura **5-1** resume las dos fases para enriquecimiento ontológico enfocada en el aprendizaje ontológico. La primera es la construcción del vocabulario a través del preprocesamiento de textos, la ponderación estadística de términos y la evaluación soportada en *gold standard*. La segunda es la extracción de conceptos y relaciones experimentando con dos técnicas, sumado a esto, las estructuras extraídas son valoradas en relación a la coherencia y su desempeño en el agrupamiento semántico.

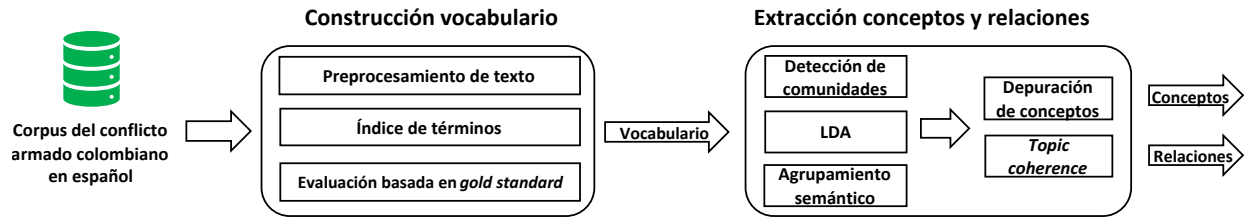

Figura **5-1**: Metodología.

5.1. Datos textuales

Para abordar el primer objetivo específico (ver capítulo 4), este estudio recuperó manualmente un conjunto de textos que aborda el conflicto armado colombiano. Además, el tesauro del CNMH [Espinosa, 2018] fue formalizado al lenguaje OWL con el propósito de utilizar la terminología, así mismo, dejar este insumo para futuras investigaciones. El capítulo 6 divulga cómo se recuperaron los textos y se formalizó el tesauro. Las siguientes secciones plantean la metodología para responder al segundo y tercer objetivo específico.

5.2. Construcción del vocabulario

Para la construcción del vocabulario se consideraron las alternativas de escritura y colocaciones. Las alternativas tomaron del tesauro del CNMH de Colombia [Espinosa, 2018]. El reconocimiento de entidades nombradas (*named entity recognition*) facilito identificar las colocaciones, es decir, se evaluó la significancia de las entidades compuestas por más de un término a través de una prueba de verosimilitud.

Este trabajo identificó automáticamente las etiquetas *part-of-speech* (POS) de los términos y filtró las palabras cuya información sintáctica fuera diferente a adjetivos, sustantivos y verbos. Las etiquetas nombradas fueron seleccionadas porque es usual asociarlas a términos informativos y útiles para la formación de conceptos [Ochoa et al., 2013]. Después, el texto sufrió el proceso de lematización y normalización a minúscula. Adicionalmente, se eliminaron stopwords, signos de puntuación, caracteres numéricos y especiales. Los términos con alternativas de escritura fueron analizados, dejando la descripción más informativa; por ejemplo, el término *Derecho Internacional Humanitario* reemplazó *DIH*. Finalmente, se tokenizó el texto considerando las colocaciones y unigramas.

Este trabajo seleccionó el índice de términos al implementar y comparar seis esquemas de ponderación, esto con el propósito de construir un vocabulario de términos relevantes para describir el contenido de los textos. Los esquemas escogidos provienen de investigaciones semejantes [Ochoa et al., 2013, Galicia-Haro and Gelbukh, 2014, Meijer et al., 2014]. A continuación, se explica cada uno:

- **TF-IDF (Term frequency-inverse document frequency):** Cuantifica la frecuencia relativa de los términos en un documento en comparación con la proporción inversa de ese término en el corpus. Este trabajo empleó la TF normalizada por la longitud de cada documento. La Eq. 5-1 presenta la ponderación TF-IDF para el término t considerando que D es el conjunto de documentos que forman el corpus, $|D|$ es la longitud del corpus, $|d_j|$ es el largo del documento j, $f_j(t)$ es la frecuencia del término t en el texto j y *f(t)* es la frecuencia de t en el corpus.

$$\text{TF-IDF}_t = \sum_{d_j \in D} \frac{f_j(t)}{|d_j|} \left(\log \frac{|D|}{f(t)} \right) \tag{5-1}$$

- **TF–Entropía:** Expresa la frecuencia relativa de los términos en un documento y el número de veces que el término aparece en cada texto analizado. Para evitar indeterminaciones matemáticas cuando un término no aparece en un documento, se aproximó a cero el logaritmo de la frecuencia del término en el texto y el corpus. La Eq. 5-2 expone el esquema de ponderación siguiendo la nomenclatura de la Eq. 5-1.

$$\text{TF-Entropía}_t = \sum_{d_j \in D} \frac{f_j(t)}{|d_j|} \left(1 + \left(\frac{1}{\log |D|} \sum_{d_j \in D} \frac{f_j(t)}{f(t)} \log \frac{f_j(t)}{f(t)} \right) \right) \tag{5-2}$$

- **Estudio de Ochoa et al. [Ochoa et al., 2013]:** Este estudio calcula la TF para todos los términos y diferencia el esquema global de ponderación. Para unigramas, existe una cuantificación de IDF, y para las colocaciones se considera una medición de los esquemas NC-value y C-Value.

  El NC-Value identifica los adjetivos, verbos y sustantivos que conforman la vecindad de un término candidato, siendo esta vecindad la historia de 10 palabras. La métrica utiliza esta información sintáctica para calcular un factor de ponderación que asigna un valor alto a colocaciones rodeadas por palabras con las categorías de interés.

  La ponderación C-Value está en Eq. 5-3 donde t es el término candidato, $|t|$ es la longitud del término candidato, *f(t)* es la frecuencia del término candidato en el corpus, T_t es el conjunto

de términos que contienen a t (*i.e.* conjunto de palabras que poseen como sub-término a t), $P(T_t)$ es la frecuencia en el corpus de la palabra más larga conformada por el término candidato y $\sum_{b_t \in T_t} f(b_t)$ es la frecuencia de aparición del término candidato como sub-término de cualquier palabra b que pertenece a T_t. C-Value asigna un valor alto a los términos que aparecen frecuentemente en el corpus y escasamente dentro de otros términos.

$$\text{C-Value}_t = \begin{cases} \log_2 |t| f(t) & \text{si } T_t \in \emptyset \\ \log_2 |t| \left(f(t) - \frac{1}{P(T_t)} \sum_{b_t \in T_t} f(b_t) \right) & \text{si } T_t \notin \emptyset \end{cases} \tag{5-3}$$

- **Modificación al estudio de Ochoa et al. [Ochoa et al., 2013]:** Este trabajo propone modificar el estudio de [Ochoa et al., 2013]. En la Eq. 5-4 se especifica la ponderación para unigramas y colocaciones. La reforma es cuantificar los términos mediante el esquema de entropía con el fin de obtener un vocabulario más preciso.

$$\begin{aligned} \text{Unigrama}_t &= \text{TF-Entropía}_t \\ \text{Colocación}_t &= \text{TF}_t \left(\text{Entropía}_t + \text{NC-Value}_t + \text{C-Value}_t \right) \end{aligned} \tag{5-4}$$

- **Estudio de Meijer et al. [Meijer et al., 2014]:** La relevancia de un término es la suma entre *domain pertinence*, *lexical coherence*, *domain consensus* y la entropía. La primera medida favorece términos frecuentes en el corpus analizados y con baja aparición en un corpus contrastivo, así recuperar términos representativos de un dominio. La Eq. 5-5 señala *domain pertinence*, siendo $f(t)_D$ la frecuencia del término t en el corpus analizado y $f(t)_{Dj}$ la frecuencia en el corpus contrastivo. Este trabajo utilizó como corpus contrastivo el Wikicorpus v. 1.0 en español [Reese et al., 2010].

$$\text{Domain pertinence}_t = \frac{f(t)_D}{f(t)_{Dj}} \tag{5-5}$$

Lexical coherence determina qué tan bien un término compuesto es representado por los términos individuales que lo componen, de tal forma que las colocaciones frecuentes, en comparación con sus términos individuales, obtiene un puntaje alto. La Eq. 5-6 presenta esta métrica donde w es un término que compone a la colocación t, $|t|$ es la longitud de t y *f(t)* es la frecuencia del término t en el corpus analizado. *Domain consensus* penaliza términos muy frecuentes dentro del corpus. Este esquema de ponderación está en la Eq. 5-7 donde D es el corpus de trabajo y $f_j(t)$ es la frecuencia del término t en el documento j, cuando el término no estaba en un documento se aproxima a cero la expresión $\log f_j(t)$.

$$\text{Lexical coherence}_t = \frac{|t| f(t) \log f(t)}{\sum_{w \in t} f(w)} \tag{5-6}$$

$$\text{Domain consensus}_t = -\sum_{d_j \in D} f_j(t) \log f_j(t) \tag{5-7}$$

Cuadro **5-1**: Datos para ejemplificar modificación al estudio de [Meijer et al., 2014]

Término	**Frecuencia f(t)**	**Término**	**Frecuencia f(t)**
violencias_sexual	502	violencia_sexual_conflicto	42
violencia	416	**junto_accion_comunal**	36
sexual	63	junto	64
reducir_violencia_sexual	7	accion	155
victima_violencia_sexual_mujer	3	comunal	11

*Los términos están en el corpus analizado y surgen del preprocesamiento donde se efectuó lematización y eliminación de signos diacríticos. Las frecuencias son los valores verídicos de la ocurrencia de los términos en el corpus.

- **Modificación al estudio de Meijer et al. [Meijer et al., 2014]:** La Eq. 5-8 presenta el esquema de ponderación para unigramas y colocaciones que surge de modificar el estudio de [Meijer et al., 2014]. En particular, este trabajo propone emplear la métrica C-Value en contraposición a *lexical coherence* para ponderar colocaciones.

$$\begin{aligned} \text{Unigrama}_t &= \text{Domain pertinence}_t + \text{Domain consensus}_t + \text{Entropía}_t \\ \text{Colocación}_t &= \text{Domain pertinence}_t + \text{C-Value}_t + \text{Domain consensus}_t + \text{Entropía}_t \end{aligned} \tag{5-8}$$

La modificación planteada no penaliza colocaciones relevantes que están compuestas por unigramas muy frecuentes. Para ejemplificar el impacto positivo que tiene utilizar C-Value en lugar de *lexical coherence*, a continuación, se presenta estos esquemas de ponderación para los términos **violencias_sexual** y **junto_accion_comunal** que hacen parte del corpus analizado.

	violencia_sexual	**junto_accion_comunal**
C-Value	$\log_2 2\left(502 - \frac{1}{4}(7+3+42)\right) = 9{,}9334$	$\log_2 3 * 36 = 6{,}169$
Lexical coherence	$\frac{2*502*\log 502}{416+63} = 5.661$	$\frac{3*36*\log 36}{64+155+11} = 0{,}7308$

*Los términos están en el corpus analizado y surgen del preprocesamiento donde se efectuó lematización y eliminación de signos diacríticos.

Los ejemplos presentados emplean los datos expuesto en el Cuadro **5-1** donde esta la ocurrencia de los términos analizados, los unigramas que conforman estas palabras y (cuando existen) colocaciones que contienen las unidades examinadas.

Estos casos demuestran que términos como **violencias_sexual** obtienen una puntuación menor bajo el esquema *lexical coherence* en relación con la métrica C-Value. Lo anterior ocurre porque **violencias_sexual** está conformada por unigramas ocurrentes como *violencia*, por lo anterior, *lexical coherence* es un esquema sesgado ante colocaciones construidas con unigramas frecuentes. Esta situación es aún más notoria al estudiar **junto_accion_comunal** pues *junto* y *accion* son lemmas con mayor ocurrencia que su respectiva colocación, por ende, el término obtiene una ponderación con una reducción de 88.15 % al comparar el valor que obtiene en C-Value.

Para seleccionar el índice de términos, este trabajo utilizó la evaluación basada en *gold standard* por dos razones. Primera, los resultados son reproducibles y comparables al examinar el mismo corpus [Konys, 2019]. Segunda, esta validación facilita la adquisición de métricas como *precision*, *recall*, y *F-measure* que caracterizan la funcionalidad de la ontología aprendida a través de los términos que son los bloques de construcción iniciales.

El inconveniente de emplear un *gold standard* fue el costo asociado a su elaboración, no obstante, sólo se incurrió en estos una vez ya que las evaluaciones posteriores fueron completamente automáticas [Dellschaft and Staab, 2008]. Para la construcción de la referencia, este trabajo consideró términos relevantes que explícitamente aparecían en el corpus y no los conceptos del dominio que podían o no estar en los textos. Por ejemplo, el *gold standard* tiene los términos *banda crimen*, *banda crimen organizado* y *banda criminal* aún cuando estas unidades pueden ser representadas por el concepto *crimen organizado* dada su similitud[1].

5.3. Extracción de conceptos y relaciones

Esta investigación se aleja de estudios como [Ochoa et al., 2013, Galicia-Haro and Gelbukh, 2014] que consideran la ponderación estadística para extraer conceptos. Además, este trabajo captura relaciones taxonómicas simultáneamente a la formación de conceptos. Las relaciones son de subsunción ya que surgen de la co-ocurrencia entre los términos.

Este trabajo propone la validación de conceptos y relaciones al analizar la coherencia de las estructuras conceptuales y examinar el desempeño de las entidades extraídas en la tarea de agrupamiento semántico de los documentos [Ali and Melton, 2018]. Este tipo de agrupación surgió para responder a dos deficiencias de los comunes algoritmos de agrupamiento soportados en el modelo de espacio vectorial. La primera es la producción de conglomerados disímiles y con bajos niveles de precisión por la alta dimensionalidad de los datos [Liu et al., 2013]. La segunda es ignorar las relaciones semánticas inherentes que vinculan a los documentos entre sí, al considerar que la única característica de clasificación es la frecuencia de aparición de los términos [Ali and Melton, 2018]. En este sentido, el agrupamiento semántico tiene como objetivo agrupar los textos en tópicos o conceptos considerando las relaciones implícitas en el corpus.

La tarea propuesta cumple con los dos criterios indicados por [Dellschaft and Staab, 2008] (ver sección 3.5) ya que este trabajo plantea usar métricas que cuantifican automáticamente los resultados de agrupamiento en función de la densidad, superposición y similitud con datos de referencia, además, la coherencia de los conceptos identificados es evaluada automáticamente. Por lo anterior, la propuesta facilita la valoración frecuente y a gran escala. Adicionalmente, la tarea es independiente del dominio.

La siguiente sección describe la extracción de estructuras ontológicas, además las métricas usadas durante la evaluación son presentadas.

5.3.1. Extracción de estructuras ontológicas

Esta investigación implementó dos escenarios para la extracción de conceptos y relaciones. El primero emplea los modelos generativos siendo el más empleado el modelo Latent Dirichlet Allocation (LDA) [Blei et al., 2003], mientras que el segundo, apunta a los algoritmos para la detección de comunidades [Fortunato, 2010].

LDA es usado ampliamente durante la recuperación de información y ha sido clasificado como una técnica para el agrupamiento automático ya que halla tópicos de un corpus y asigna distribuciones de estos a cada documento, además, determina distribuciones de términos sobre los tópicos [Blei et al., 2003]. Cada tópico construido es un concepto pues es una agrupación de términos que poseen una alta probabilidad de pertenecer a este. El número de tópicos T a identificar dentro de los textos es un parámetro calibrado siguiendo alguna especificación (por ejemplo, perplejidad) [Blei

[1] El concepto está en el tesauro del CNMH de Colombia [Espinosa, 2018].

et al., 2003]. El modelo LDA forma conceptos empleando únicamente el corpus, en consecuencia, este trabajo se encausa en el enfoque basado en datos asociado al aprendizaje ontológico (ver sección 3.4).

LDA asume que los documentos pueden ser descritos como una combinación probabilística de los T tópicos, en este sentido, existe un vector θ_d que describe la probabilidad de que el documento d pertenezca a cada uno de los T tópicos [Edison and Carcel, 2021]. El vector θ_d tiene la forma descrita en Eq. 5-9 y la matriz θ (con dimensiones DxT) describe al corpus.

$$\theta_d = \begin{bmatrix} P(t_1|d) & P(t_2|d) & P(t_3|d) & ... & P(t_{T-1}|d) & P(t_T|d) \end{bmatrix} \tag{5-9}$$

Igualmente, cada $t \in T$ es caracterizado por una distribución de probabilidad sobre el vocabulario de tamaño V, por lo tanto, identificar un término depende de la probabilidad condicional asociada a un tópico. Por ejemplo, detectar términos como *farc, eln, guerrilla* dependerá de la presencia del tópico *grupos guerrilleros*[2] dentro de los textos. La Eq. 5-10 describe la probabilidad del vocabulario sobre el tópico t. El vector φ_t permite conocer la composición del tópico ya que los términos con mayor probabilidad indican el contenido de t. La matriz φ (con dimensiones VxT) especifica a los T tópicos.

$$\varphi_t^T = \begin{bmatrix} P(w_1|t) & P(w_2|t) & P(w_3|t) & ... & P(w_{V-1}|t) & P(w_V|t) \end{bmatrix} \tag{5-10}$$

Con la información θ y φ, el modelo LDA [Blei et al., 2003] plantea la creación de texto al, primero, establecer una distribución de probabilidad de los términos siguiendo $\varphi \sim Dir(\beta)$, segundo, determinar una probabilidad de los tópicos sobre los documentos tal que $\theta \sim Dir(\alpha)$, tercero para el conjunto de términos $N_d = (w_{d,1}, ...w_{d,n})$ dentro del documento d, asignar un tópico de manera que $z_{d,n} \sim Mult(\theta_d)$ y disponer cada $w_{d,n}$ mediante $p\left(w_{d,n}|z_{d,n}, \varphi\right)$. En particular, β y α son hiperparámetros necesario para el muestro de Gibbs que es la técnica a través del cual se actualizan las asignaciones de tópicos de los términos condicionados a la adjudicación de tópicos a todo el vocabulario [Edison and Carcel, 2021]. La Eq. 5-11 expresa la probabilidad del texto en relación a los parámetros del modelo.

$$\prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \left(\prod_{n=1}^{N_d} \sum_{z_{d,n}} p\left(z_{d,n}|\theta_d\right) p\left(w_{d,n}|z_{d,n}, \varphi\right) \right) \tag{5-11}$$

En Eq. 5-11 la expresión $p\left(\theta_d|\alpha\right)$ describe la probabilidad de observar la distribución de tópicos sobre θ_d, además $p\left(z_{d,n}|\theta_d\right)$ determina cuán probable es la asignación individual del tópico $z_{d,n}$ (relacionado al término n del documento d) que está condicionada a la distribución de tópicos sobre el documento (θ_d). Así mismo, $p\left(w_{d,n}|z_{d,n}, \varphi\right)$ es la probabilidad de detectar una palabra $w_{d,n}$ condicionada por la asignación del tópico sobre la palabra ($z_{d,n}$) y las probabilidades de palabras sobre los tópicos (φ).

La Eq. 5-11 establece la probabilidad de observar los textos en el corpus mediante la suma sobre las posibles asignaciones de tópicos, el producto entre todos los términos del documento d y el producto en relación a todos los textos del corpus [Edison and Carcel, 2021].

[2]Este ejemplo fue construido con información que está en el tesauro del CNMH [Espinosa, 2018]

Este trabajo utilizó *Topic Coherence* (TC) para calibrar el modelo LDA. Esta es métrica de la interpretabilidad semántica de cada tópico descubierto [Korenčić et al., 2018]. Este trabajo empleó la métrica de [Mimno et al., 2011] presentada en Eq. 5-12.

$$\mathrm{TC}\left(t; V^{(t)}\right) = \frac{2}{N(N-1)} \sum_{m=2}^{N} \sum_{l=1}^{m-1} \log \frac{f\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{f\left(v_l^{(t)}\right)} \tag{5-12}$$

Donde $V^{(t)} = (v_1^{(t)}, ..., v_N^{(t)})$ es la lista de los N top términos dentro del concepto (o tópico) t, es decir, los términos más probables dentro de la temática. *f(v)* es la frecuencia del término v en el corpus, y $f(v, v^{'})$ es el número de textos que contienen simultáneamente a v y $v^{'}$.

Los algoritmos de detección de comunidades parten de la matriz de co-ocurrencia para establecer una red de palabras. Estos algoritmos agrupan nodos (términos) que comparten propiedades comunes o poseen roles similares dentro del grafo [Fortunato, 2010]. En este sentido, conceptos que describen el contenido del corpus son establecidos, y después, los documentos asociados semánticamente son conglomerados en en cada comunidad o concepto establecido [Liu et al., 2013].

Este trabajo empleó una red dirigida de palabras ponderadas para explotar la información semántica que surge de la aparición ordenada de los términos. La Figura **5-2** presenta un ejemplo de la red donde el peso del arco entre el término t_i y t_j, siendo $i \neq j$, existía si la co-ocurrencia entre las palabras era significativa bajo una prueba de verosimilitud con 95 % de confianza.

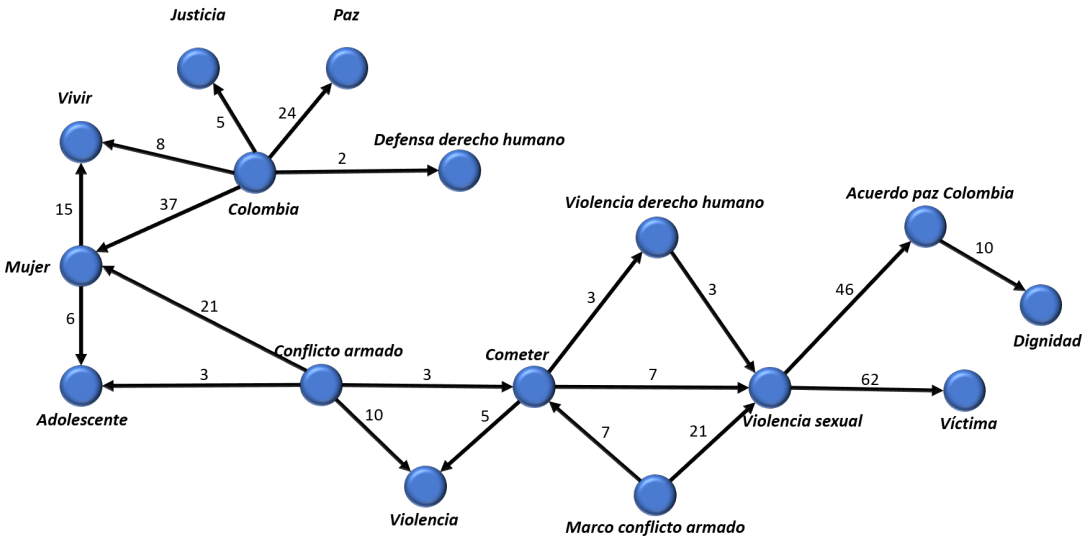

Figura **5-2**: Ejemplo de red dirigida de palabras ponderadas

Este trabajo empleó el algoritmo *Directed Louvain* [Dugué and Perez, 2015] que despliega una estructura jerárquica de las comunidades empleando la noción de modularidad dirigida presentada en [Leicht and Newman, 2008]. El empleo de *Directed Louvain* en redes asimétricas ha demostrado construir comunidades precisas ya que examina el desbalanceo en los grados de entrada y salida, lo cual es una característica de las redes de palabras. Este algoritmo ha sido criticado por la diferencia en los resultados debido al orden con que se analizan los nodos durante la agrupación. Por esto, este trabajo empleó el agrupamiento consensuado (*consensus clustering*). En específico, se usó la propuesta de [Lancichinetti and Fortunato, 2012] quienes plantean una matriz de consenso donde d_{ij} indica el número de particiones en las que los vértices i y j fueron asignados al mismo grupo. Los

pesos de la matriz de consenso se filtraron si no eran significativos con una prueba de verosimilitud al 95 % de confianza.

Después de la detección de comunidades, el siguiente paso fue el agrupamiento semántico de los documentos en cada comunidad. Este estudio empleó *Pointwise Mutual Information* (PMI) para medir el grado de asociación entre documentos y conceptos. Esta estrategia es similar a la reportada en [Liu et al., 2013]. La Eq. 5-13 presenta PMI entre el documento d_k y el concepto C_l.

$$\mathrm{PMI}\left(d_k, C_l\right) = \frac{\log\left(p\left(d_k\right) p\left(C_l\right)\right)}{\log\left(p\left(d_k, C_l\right)\right)} \tag{5-13}$$

Donde $p(d_k, C_l)$ es la probabilidad del documento d_k, y el concepto C_l. Esto proviene de la similitud coseno como está en la Eq. 5-14.

$$p\left(d_k, C_l\right) = \frac{\sum_{i=1}^{V} w\left(i, d_k\right) w\left(i, C_l\right)}{\sqrt{\sum_{i=1}^{V} w^2\left(i, d_k\right) \sum_{i=1}^{V} w^2\left(i, C_l\right)}} \tag{5-14}$$

Donde V es el tamaño del vocabulario, $w(i, d_k)$ es el peso del término i en el documento d_k utilizando la información de la *matriz término – documento* con ponderación TF. $w(i, C_l)$ es el peso del término i en el concepto C_l. La matriz $P = w(i, C_l)$ proviene del algoritmo *Weighted Leader Rank* (WLR) [Lü et al., 2011] que pondera la importancia de un nodo (término) en la red (concepto).

Esta investigación utilizó el algoritmo WLR reportado en [Xuan et al., 2012] que genera un puntaje S_i para cada nodo i al realizar una modificación en la matriz P y usufructuando el *out-degree* de los vértices. La modificación consiste en generar un grafo conectado mediante la adición de un nodo virtual σ, de forma que exista un enlace bi-direccional entre cada nodo i y σ. El peso de los arcos bi-direccionales (*i.e.* $w_{i\sigma}$ y $w_{\sigma i}$) es ajustado a uno. El nodo virtual con enlaces bi-direccionales permite que la matriz P sea primitiva[3] [Lü et al., 2011]. Igualmente, WLR utiliza el *out-degree* del nodo i (*i.e.* o_i) para reflejar la importancia del vértice respecto a la información que transmite.

WLR plantea el rankeo de nodos como cambios en una serie de tiempo donde $s_i(t)$ es el puntaje del nodo i en el tiempo t. La Eq. 5-15 describe $s_i(t)$ considerando a V el número de nodos en el grafo (*i.e.* tamaño del vocabulario), además el puntaje inicial de los vértices i es $s_i(0) = 1$ y $s_\sigma(0) = 0$.

$$s_i(t) = \sum_{j=0}^{V} \frac{w_{ij} s_j(t-1)}{o_j} \tag{5-15}$$

Dado que P es una matriz primitiva que converge en un tiempo finito t_c[4], el puntaje total S_i se establece en la Eq. 5-16 donde M es un parámetro de normalización descrito por $M = \frac{s_\sigma(t_c)}{V + \max s_i(t_c)}$,

[3] Matriz cuadrada con todos sus valores positivos y un único vector propio con valor propio igual a uno, por ende, P^T tiene un único estado estable. En [Lü et al., 2011] demuestran que P converge a un único estado estable en un tiempo finito. La demostración mencionada está como suplemento del artículo de [Lü et al., 2011], este material es de libre acceso en `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021202#s5`

[4] En [Lü et al., 2011] plantean controlar el tiempo de convergencia a través de la diferencia entre el puntaje del tiempo anterior y el actual, es decir, la difusión al estado estable se detiene cuando $|s_i(t) - s_i(t-1)| < umbral$ considerando $umbral = 0{,}00002$. Esta investigación implementó este criterio de parada.

de esta forma, $0 < S_i \leq 1$.

$$S_i = \frac{\left(\frac{s_i(t_c)+s_\sigma(t_c)}{V}\right)}{M} \tag{5-16}$$

Es así como WLR es muy semejante a *PageRank*, no obstante, WLR es un algoritmo adaptativo y no paramétrico, pues en cada iteración establece los puntajes considerando información de pasos previos. Además, WLR cuantifica la relevancia de un nodo a partir del número y peso de sus vecinos. En este sentido, WLR supera a *PageRank* en la robustez para manipular datos ruidosos produciendo mayor precisión al ponderar la importancia de los nodos [Lü et al., 2011].

Continuando la descripción de los elementos asociados a la Eq. 5-13, se plantea que la probabilidad del documento d_k es:

$$p\left(d_k\right) = \frac{\sum_{i=1}^{V} w\left(i, d_k\right)}{\sum_{j=1}^{D}\sum_{i=1}^{V} w\left(i, d_j\right)} \tag{5-17}$$

Recordando que el tamaño del vocabulario es V y el número de documentos es D. Adicionalmente, $p(C_l)$ es la probabilidad del concepto C_l como está en la Eq. 5-18, donde L es el número de conceptos.

$$p\left(C_l\right) = \frac{\sum_{i=1}^{V} w\left(i, C_l\right)}{\sum_{j=1}^{L}\sum_{i=1}^{V} w\left(i, C_j\right)} \tag{5-18}$$

PMI estableció la asociación entre el documento d_k y el concepto C_l en un rango de [-1,1], siendo -1 que estos objetos nunca co-ocurren simultáneamente, 0 que son independientes y 1 que siempre co-ocurren [Amoualian et al., 2017]. Este estudio fijó que todas las asociaciones negativas pasaran a ser cero y los demás valores quedaron igual. Esta estrategia ha sido empleada en [Levy et al., 2015] para la representación truncada de las palabras, de esta forma, abordar la poca fiabilidad de los valores negativos de PMI en corpus pequeños.

Esta estrategia permitió generar una matriz $Ass = PMI(d_k, C_l)$ de grados de asociación entre documentos y conceptos, en particular, las dimensiones de Ass son DxL. En consecuencia, este trabajo detectó una *possibilistic partition* de los documentos en función de los conceptos detectados. Es muy importante subrayar que el grado de asociación entre el documento d_k y el concepto C_l es una posibilidad, por ende, nunca debe ser interpretada como una probabilidad.

5.3.2. Métricas para evaluar las estructuras extraídas

Este trabajo consideró métricas para: i) evaluar el desempeño de los conceptos en la tarea de agrupamiento semántico y ii) cuantificar automáticamente la coherencia de las estructuras conceptuales. Se emplearon índices *soft partition* al producirse una *possibilistic partition*.

- **Índice Dunn:** Medida de densidad más recomendada y sencilla de calcular para identificar grupos separados y densamente distribuidos. Este trabajo calculó la densidad *Dunn* de la

agrupación como la suma de los índices de densidad en cada partición d. La Eq. 5-19 presenta la definición matemática del índice Dunn.

$$Dunn = \sum_{l=1}^{L} \frac{1}{D} \sum_{k=1}^{D} den(d_k, C_l) \qquad den(d_k, C_l) = \begin{cases} g(d_k, C_l) & si\ g(d_k, C_l) \geq T_c \\ 0 & si\ g(d_k, C_l) < T_c \end{cases} \tag{5-19}$$

En la Eq. 5-19 el valor $g(d_k, C_l)$ referencia la información de la matriz de probabilidad para el modelo LDA y la matriz Ass (*i.e.* matriz de grados de asociación) para *Directed Louvain*. Este trabajo estableció $T_c = 0{,}6$. El índice Dunn arroja un valor cercano a cero cuando un documento tiene igual grado de asociación con los conceptos donde se pretende agrupar.

- **Medida de superposición:** Evalúa la superposición de los grupos al considerar la información de la matriz de probabilidad y la matriz de grados de asociación. Esta investigación empleó la propuesta de [Lin et al., 2016] que está descrita en las Eq. 5-20 y Eq. 5-21.

$$Superposición = \frac{2}{L(L-1)} \sum_{C_p \neq C_q} \sum_{d_k \in C_p \cup C_q} h\left(d_k, C_p, C_q\right) \tag{5-20}$$

$$h\left(d_k, C_p, C_q\right) = \begin{cases} \exp^{1-|g(d_k,C_p)-g(d_k,C_q)|} & si\ |g(d_k, C_p) - g(d_k, C_q)| \leq H \\ 0 & si\ |g(d_k, C_p) - g(d_k, C_q)| > H \end{cases} \tag{5-21}$$

Esta métrica establece un umbral H de 0.4 para determinar si un documento está en el área de superposición de dos particiones. Entre mayor sea este índice, las agrupaciones de documentos estarán más superpuestas.

- **Índice frand ajustado:** Medida de la familia de *fuzzy rand index* que compara la similitud entre los documentos agrupados automáticamente y una referencia manual de dicho agrupamiento. Las ventajas y problemas de emplear un listado de referencia para comparar el agrupamiento semántico son similares a las mencionadas previamente al analizar la evaluación basada en *gold standard* (ver sección 3.5). Esta investigación aplicó la propuesta reportada en [Horta and Campello, 2015] porque cumple cuatro criterios prácticos: i) alcanza su valor máximo cuando dos respuestas equivalentes son comparadas, ii) detecta la mejor solución dentro de un conjunto de soluciones, iii) muestra progresivamente mejores evaluaciones para una solución superior, y iv) está corregida garantizando que el resultado no sea producido por las fluctuaciones de aleatoriedad inherentes a la medición.

  En [Horta and Campello, 2015] construyen dos matrices. La primera es la matriz de membresía (*membership matrix*) $U = u_{ij}$ donde u_{ij} es el grado de asociación del documento i al conglomerado j. La segunda matriz $V = v_{ij}$ establece la asociación entre el texto i y la partición j utilizando el *gold standard*.

  Estas matrices permiten computar la información entre las parejas de textos, en este sentido, los autores en [Horta and Campello, 2015] reportan calcular las matrices $J^U = UU^T$ y $S^U = U(1_k - I)U^T$ donde 1_k es una matriz con 1 en cada celda. Por consiguiente, J^U y S^U son matrices cuadradas cuyo número de filas es igual a la cantidad de textos analizados. En particular, J^U_{ij} indica la probabilidad de que los documentos d_i y d_j pertenezcan al mismo conglomerado de acuerdo con U, por otro lado, S^U_{ij} da la información opuesta. Así mismo, J^V y S^V son las correspondientes matrices considerando la información de V.

  De esta forma, se define $\dot{a}$ (ver Eq. 5-22) y $\dot{d}$ (ver Eq. 5-23) como los indicadores de concordancia entre U y V respecto a la veracidad y falsedad de la afirmación "d_i y d_j pertenecen

a la misma partición".

$$\dot{a} = \sum_{ij} \text{mín}\left\{J^U_{i,j}, J^V_{i,j}\right\} \tag{5-22}$$

$$\dot{d} = \sum_{ij} \text{mín}\left\{S^U_{i,j}, S^V_{i,j}\right\} \tag{5-23}$$

Además, $\dot{b}$ (ver Eq. 5-24) y $\dot{c}$ (ver Eq. 5-25) establecen el desacuerdo entre U y V sobre la misma aseveración.

$$\dot{b} = \sum_{ij} \text{mín}\left\{J^U_{i,j} - \text{mín}\left\{J^U_{i,j}, J^V_{i,j}\right\}, S^V_{i,j} - \text{mín}\left\{S^U_{i,j}, S^V_{i,j}\right\}\right\} \tag{5-24}$$

$$\dot{c} = \sum_{ij} \text{mín}\left\{J^V_{i,j} - \text{mín}\left\{J^U_{i,j}, J^V_{i,j}\right\}, S^U_{i,j} - \text{mín}\left\{S^U_{i,j}, S^V_{i,j}\right\}\right\} \tag{5-25}$$

La Eq. 5-26 expone el índice frand considerando la información previa.

$$Frand\,(U,V) = \frac{\dot{a} + \dot{d}}{\dot{a} + \dot{b} + \dot{c} + \dot{d}} \tag{5-26}$$

Con el propósito de que los resultado del índice frand no sean efecto de fluctuaciones de aleatoriedad inherentes a la medida, los autores en [Horta and Campello, 2015] proponen computar la esperanza de la métrica. La Eq. 5-27 expone la esperanza del índice frand donde $E\left[\dot{a}\right]_{U,V}$ y $E[\dot{d}]_{U,V}$ son las esperanzas de los indicadores $\dot{a}$ y $\dot{d}$ respectivamente.

$$E\left[Frand\right]_{U,V} = \left(\dot{a} + \dot{b} + \dot{c} + \dot{d}\right)^{-1}\left(E\left[\dot{a}\right]_{U,V} + E[\dot{d}]_{U,V}\right) \tag{5-27}$$

La esperanza de $\dot{a}$ se computan siguiendo lo indicado en el Algoritmo 1, de forma homóloga es posible establecer $\dot{d}$ al ubicar las matrices S^U y S^V en los pasos uno y dos. Posteriormente, $E\left[Frand\right]_{U,V}$ permite calcular el índice frand ajustado. La Eq. 5-28 plantea esta medida, en particular, el denominador normaliza el resultado de tal forma que uno sea el valor máximo.

$$FrandAjust\,(U,V) = \frac{Frand\,(U,V) - E\left[Frand\right]_{U,V}}{1 - E\left[Frand\right]_{U,V}} \tag{5-28}$$

El índice frand ajustado obtiene un valor de uno cuando la agrupación manual y la obtenida automáticamente son iguales.

Este estudio utilizó la propuesta de [Mimno et al., 2011] para cuantificar automáticamente la coherencia de los conceptos extraídos, además se usó WLR con el propósito de identificar los N términos más relevantes dentro de cada concepto extraído con el algoritmo *Directed Louvain*. La métrica TC y el algoritmo WLR han sido explicados en la sección 5.3.1.

Algoritmo 1: Computo de $\mathrm{E}[\dot{a}]_{U,V}$

Representar el triángulo superior de J^U en el vector x;
Representar el triángulo superior de J^V en el vector y;
m$\leftarrow \frac{n(n-1)}{2}$ (Tamaño de los vectores x, y)
$E\left[\dot{a}\right]_{U,V} \leftarrow 0$
$i, j \leftarrow m, m$
while $i > 0$ **do**
    while $j > 0$ *and* $x_i \leq y_j$ **do**
        $j \leftarrow j - 1$
    end
    $\mathrm{E}[\dot{a}]_{U,V} \leftarrow E\left[\dot{a}\right]_{U,V} + x_i(m - j)$
    $i \leftarrow i - 1$
end
$i, j \leftarrow m, m$ **while** $j > 0$ **do**
    while $i > 0$ *and* $x_i > y_j$ **do**
        $i \leftarrow i - 1$
    end
    $\mathrm{E}[\dot{a}]_{U,V} \leftarrow E\left[\dot{a}\right]_{U,V} + y_j(m - j)$
    $j \leftarrow j - 1$
end
$\mathrm{E}[\dot{a}]_{U,V} \leftarrow \frac{E[\dot{a}]_{U,V}}{m}$

5.3.3. Depuración de conceptos

Las técnicas documentadas no emplean conocimiento externo para la formación de conceptos y relaciones, por lo anterior, las estructuras generadas pueden contener términos incorrectos a la luz de la interpretabilidad humana. En este orden de ideas, este trabajo propone depurar las entidades ontológicas construidas siguiendo un enfoque práctico donde no se requiera expertos humanos.

TC es una métrica que aumenta por la participación de los top términos en cada concepto [Chang et al., 2009], es decir, las últimas palabras dentro de la estructura realizan un aporte menor a la coherencia con relación a los primeros términos. En este sentido, la depuración propuesta consiste en retirar los últimos términos cuya presencia no aumenta la coherencia. En especifico, cada concepto fue analizado de la siguiente forma:

1. Se calculó la coherencia total del concepto (TC_i).
2. Se calculó la variación de la coherencia por el ingreso de cada término que perteneciera al concepto, consideran el orden de importancia de los términos según la métrica WLR (para las comunidades *Directed Louvain*) o la probabilidad del término en el tópico (para las temáticas extraídas con LDA).
3. Se estableció la media y desviación de las variaciones.
4. Se eliminaron los términos que redujeran la coherencia en un valor menor a la media menos una desviación estándar.
5. Se cuantificó la coherencia total del concepto reducido (TC_{i+1}).

6. Los pasos dos a cuatro se repitieron hasta que $TC_{i+1} - TC_i < d$ siendo d un parámetro que limita el largo de los conceptos. Conceptos con pocas palabras son interpretables por los humanos, en particular [Chang et al., 2009] indican que un número de 5 términos por tópico permite a los sujetos reconocer conceptos coherentes; no obstante, estas estructuras no son apropiadas para el agrupamiento de documentos.

Así mismo, se realizó la prueba *word intrusion* para valorar la interpretabilidad que humanos no expertos adjudican a las estructuras generada. Esta prueba se ejecutó sin importar el valor de d pues la propuesta de depuración no modifica los términos top dentro de los conceptos. Esta investigación realizó la prueba *word intrusion* siguiendo la metodología propuesta en [Chang et al., 2009], donde a cada concepto es agregado aleatoriamente un término relevante de otra unidad conceptual, después, humanos no experto analizan los listados y seleccionan la palabra que a su juicio es la intrusa en cada conjunto. Esta tarea cuantifica la precisión del modelo comprendida como el nivel de correspondencia entre los términos intrusos detectados por los sujetos y los ingresados adrede dentro de cada concepto, por lo anterior, existe un medida de precisión para cada estructura conceptual evaluada.

La prueba *word intrusion* es sesgada cuando la palabra intrusa es de una temática similar al concepto [Senel et al., 2018], es decir, la tarea es más difícil para los seres humanos cuando la palabra agregada posee un sentido cercano al descrito en la unidad conceptual. Por ejemplo, un sujeto identificará fácilmente el término *informe* como intruso dentro del listado *(farc, eln, gobierno, guerrilla, paramilitar)* en comparación con la palabra *colombia*. En este sentido, este trabajo solicito a cada sujeto realizar tres ensayo por cada concepto, considerando que en cada prueba variaba la palabra intrusa. La Eq. 5-29 expone como se valora la precisión del concepto k donde S es el número de sujetos que ejecutó la prueba, n es el ensayo ejecutado, w_k es el índice de la palabra intrusa añadida en la unidad conceptual k, $i_{k,s}$ es el índice del término que el humano s seleccionó como intruso en el concepto k, de esta forma, $i_{k,s} = w_k$ es igual a uno cuando los índices son los mismos y cero de los contrario.

$$MP_k = \frac{\sum_s \sum_{n=1}^{3} (i_{k,s} = w_k)_n}{S} \tag{5-29}$$

Esta métrica representa la fracción de sujetos que está de acuerdo con la composición de los conceptos, por ende, valores cercanos a uno implican que la unidad conceptual está constituida por términos apropiados.

Capítulo 6

Resultados y discusión

Este trabajo deja a disposición[1] el tesauro del CNMH formalizado, el corpus analizado, el listado de alternativas y stopwords, el *gold standard*, el vocabulario construido, la referencia manual de agrupamiento, así como los conceptos identificados, depurados y validados. La implementación fue en Python 3.8. Los experimentos se ejecutaron en una máquina con un procesador AMD Ryzen 5 @ 2.2 GHz/4 y 8 GB de memoria RAM.

Este apartado presenta la información siguiendo la misma organización del capítulo 5. Así mismo, la discusión de los resultados se encuentra después de exponer los hallazgos del trabajo.

6.1. Datos textuales

El corpus estuvo conformado por 240 textos en español disponibles en la web y recuperados manualmente a través de la ecuación *"conflicto armado"* y *"Colombia"*. Los criterios para incluir los textos fueron: i) disponibilidad gratuita, ii) documento escrito en el idioma español, iii) fecha de publicación entre 2013-2018 y iv) que el texto no hubiera sido incluido previamente. El corpus estuvo compuesto por 8,461 sentencias y 12,955 tokens (antes de preprocesamiento).

La autora formalizó el tesauro del CNMH [Espinosa, 2018] con ayuda del software *Protegé*, de esta forma, los términos dentro de este instrumento fueron manipulables. Es necesario mencionar que toda la información del tesauro no fue formalizada pues algunas relaciones están escritas de forma ambigua, por ejemplo, el tesauro bajo la *relación asociativa* agrupa conceptos que se vinculan por subordinación (*i.e.* el desplazamiento forzado es un crímenes de guerra) y estructuras que se relacionan de manera causal (*i.e.* los accidentes causan lesiones)[2]; aún cuando en los lenguajes formales existe una relación particular de subsunción y otra de causa. Por lo anterior, si la autora hubiera formalizado esta información, tendría que haber realizado suposiciones sobre la naturaleza de las *relaciones asociativas*. El tesauro formalizado tiene 1158 clases y 3784 relaciones del tipo *is-a* y *part-of* ya que estas correspondencias aparecen textualmente en el tesauro.

[1] `https://github.com/madegomez/Ontology-learning-spanish`

[2] Los ejemplos señalados aparecen explícitamente en el tesauro del CNMH [Espinosa, 2018].

6.2. Construcción del vocabulario

Para construir el vocabulario, este trabajo empleó las herramientas de reconocimiento de entidades, tokenización y lematización del paquete Freeling 4.0[3]. Al establecer las colocaciones, se tomaron las entidades que poseían más de una palabra y se filtraron aquellas cuya probabilidad de ocurrencia no era significativa bajo la prueba de verosimilitud con una confianza del 95 %. El listado de stopwords empleado fue la compilación de las listas disponibles para el tratamiento del idioma español en las aplicaciones Orange3, NLTK 3.4.5 y Google code project. De esta forma, se construyó un vocabulario de 10,182 términos conformado por colocaciones y unigramas.

Este estudio construyó manualmente el *gold standard* utilizada para seleccionar el vocabulario. La referencia está formada por 5,113 términos de los cuales 2,806 son unigramas y las demás colocaciones.

Para seleccionar el conjunto de términos que eficientemente representaba los documentos, se siguió el método de filtrado reportado en [Silva and Ribeiro, 2010] donde:

1. Se organizaron ascendentemente los listados de términos adquiridos mediante los esquemas de ponderación nombrados en el apartado 5.2.
2. Se generó un nuevo listado de términos al remover la palabra en la última posición, es decir, se descartaba el término que dado el esquema de ponderación era el menos relevante.
3. Se calculaba la *F-measure* para el nuevo listado.
4. Se repitió el segundo y tercer paso hasta que no quedaron términos para representar el corpus.
5. Se comparó los resultados de la *F-measure* para todas las iteraciones y esquemas de ponderación. Así, se seleccionó el listado de términos que presentaba mayor similitud contra el *gold standard*.

El Cuadro **6-1** presenta las características de los listados que exhibían mejores *F-measure* para cada esquema. En el Cuadro **6-1**, las dos últimas columnas muestran el número de colocaciones y unigramas recuperados que aparecen en el *gold standard*. Todos los esquemas reportados empleaban la misma ponderación local del término, es decir, la TF y diferentes pesos globales.

Cuadro **6-1**: Caracteristicas de listados de términos

Esquema	**Tamaño vocabulario**	**F–measure**	**Precision**	**Recall**	**No. colocaciones relevantes**	**No. unigramas relevantes**
TF-IDF	7,935	0.651	0.535	0.829	1,777	2,469
TF-Entropía	7,997	0.652	0.534	0.835	1,785	2,488
Estudio de Ochoa et al. [2013]	8,153	0.661	0.538	0.857	1,903	2,484
Modificación a estudio de Ochoa et al. [2013]	6,810	0.676	0.592	0.787	1,731	2,300
Estudio de Meijer et al. [2014]	6,226	0.713	0.649	0.791	2,000	2,023
Modificación a estudio de Meijer et al. [2014]	6,176	0.738	0.674	0.814	2,215	1,948

La modificación al estudio de Meijer et al. [2014] recuperó el vocabulario con el mayor número de colocaciones relevantes ya que esta métrica ubica a los términos compuestos en posiciones superiores,

[3] `http://nlp.lsi.upc.edu/freeling/`

en comparación con los índices producidos bajo los otros esquemas. Así, este esquema tiene el mejor nivel de *recall*.

TF-Entropía recuperó unigramas más similares a los de la lista de referencia. Por ende, la entropía extrae unigramas más relevantes que la ponderación IDF porque considera simultáneamente la frecuencia del término en el corpus y en cada texto. El esquema TF-Entropía tiene un *recall* muy similar a la modificación del estudio de Meijer et al. [2014] dado el número de unigramas recuperados que verdaderamente pertenecían al *gold standard*.

La modificación al estudio de Ochoa et al. [2013] tiene un desempeño mejor que las ponderaciones TF-IDF y TF-Entropía, cuando la *F-measure* es examinada, porque discrimina la cuantificación de las colocaciones. Este resultado es consecuencia de la *precision* lograda al recuperar un índice de términos más pequeño y constituido por palabras relevantes.

El estudio de Meijer et al. [2014], así como su modificación, tienen un comportamiento superior a todos los otros esquemas ya que el análisis contrastivo penaliza unigramas que aún frecuentes no describen el dominio, además, esta estrategia favorece a las colocaciones. En este sentido, los índices tiene una distribución de términos muy semejante al *gold standard*. En particular, la modificación al estudio de Meijer et al. [2014] genera el vocabulario con el menor número de términos produciendo el mayor nivel de precisión y, en consecuencia, tiene el mejor desempeño de la medida *F-measure*.

La modificación del estudio de Meijer et al. [2014] tiene un valor *F-measure* similar a los reportados en [Ochoa et al., 2013, Galicia-Haro and Gelbukh, 2014], aún cuando el esquema de ponderación es diferente ya que las investigaciones nombradas no realizan análisis contrastivo, sino que usufructúan los patrones léxicos. Lo anterior podría considerarse una ventaja del trabajo aquí documentado, dado que no se requieren expertos humanos ni fuentes de conocimiento especializados para establecer la forma sintáctica que describe los términos del dominio; sin embargo, la propuesta planteada es dependiente de la distribución de términos y calidad del corpus contrastivo [Al-Aswadi et al., 2020].

Así mismo, los resultados del Cuadro **6-1** pueden ser efecto de la evaluación basada en un *gold standard* elaborado manualmente, por consiguiente, es posible que los listados de términos extraídos capturen datos no considerados por la persona que construyó el listado de referencia [Petasis et al., 2011].

El listado de términos adquirido al modificar el estudio de Meijer et al. [2014], aún con las objeciones señaladas, tiene el más alto valor *F-measure* siguiendo la metodología descrita. Por ende, este vocabulario fue empleado como insumo de entrada para la extracción de los conceptos y relaciones del corpus.

6.3. Extracción de conceptos y relaciones

Este estudio segmentó el corpus siguiendo la división clásica 70-30 para realizar la tarea de agrupamiento semántico de documentos. La división provino de un proceso aleatorio, además, el método 5-folds fue empleado en búsqueda de la validez de resultados. El conjunto de entrenamiento estuvo formado en promedio por 6,176 sentencias, y los datos de evaluación tenían en promedio 1,799 oraciones. Es necesario clarificar que los mismos conjuntos de entrenamiento fueron usados para extraer las estructuras ontológicas mediante el modelo LDA y *Directed Louvain*.

Para la calibración de LDA no fue necesario ajustar el modelo a datos previamente no vistos ya que este estudio utilizó la propuesta de TC reportada en [Mimno et al., 2011]. La aplicación de TC consideró el rango de 5 a 60 términos top dentro de cada concepto. Así el mayor valor de TC se obtuvo para 5 tópicos. El valor de la distribución previa del tema en el documento fue de 0.28 y

la distribución a priori del término en el tópico de 0.01. Esto permitió obtener un TC con media máxima de $-6{,}988 \pm 0{,}261$ para N=5 y un valor mínimo de $-9{,}032 \pm 0{,}047$ para N=60 con un 95 % de confianza. A partir del modelo construido, se agruparon los documentos del conjunto de evaluación.

Para el algoritmo *Directed Louvain*[4], el insumo inicial fue una red dirigida de palabras con una ventana igual a dos. El grafo tenía 6,176 nodos y 34,169 arcos. *Directed Louvain* entregó tres niveles jerárquicos cuyas particiones tenían una modularidad media de $0{,}3353 \pm 0{,}00085183$ con 95 % de confianza.

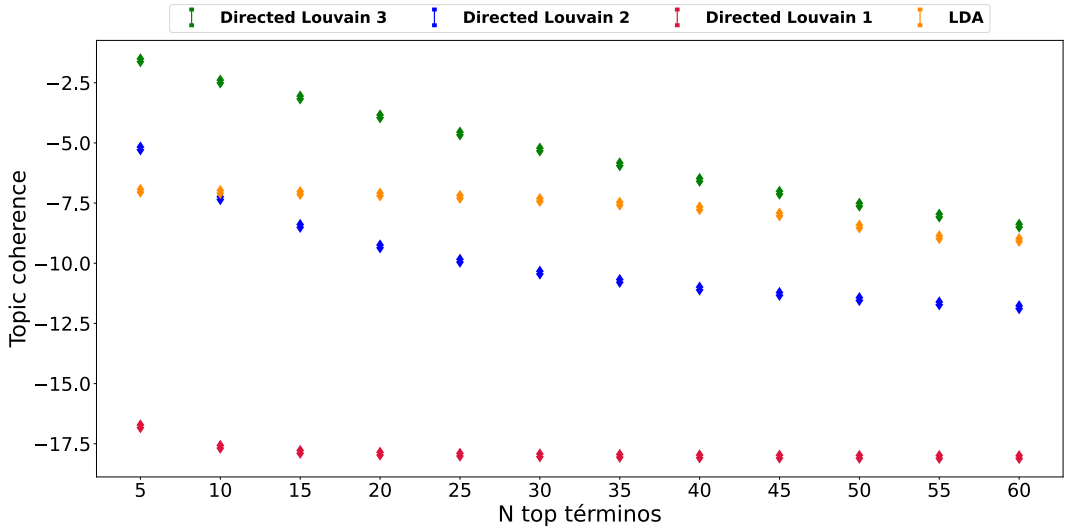

Figura **6-1**: Coherencia del Tópico (*Topic coherence* - TC) con top términos entre 5 a 60.

Esta investigación construyó los conceptos siguiendo una metodología similar a la reportada en [Qiu et al., 2020], no obstante, estos estudios tienen dos aspectos diferentes. Primero, el trabajo aquí documentado generó una red dirigida para capturar la información semántica asociada al orden de los términos, diferente a [Qiu et al., 2020] quienes emplearon una red no dirigida. Segundo, la validación de los conceptos construidos. En [Qiu et al., 2020] evaluaron la calidad de los conceptos mediante la modularidad y comparación contra un *gold standard*. La evaluación guiada por la modularidad favorece a los conceptos construidos utilizando algoritmos centrados en la optimización de esta métrica, por lo anterior, la evaluación es sesgada. Por ejemplo, si este trabajo hubiera utilizado este marco, habría perjudicado a los conceptos construidos por LDA en comparación con los de *Directed Louvain*. Adicionalmente, la evaluación basada en *gold standard* no se pudo aplicar en este estudio porque esta referencia conceptual no existe en el dominio investigado, y el objetivo era evitar su construcción ya que implicaba altos costos.

La Figura **6-1** expone el TC medio y la barra de error de los escenarios experimentales cuando los términos top varían de 5 a 60. Para *Directed Louvain* se indica el nivel de jerarquización con el número después del nombre, siendo 1 el primer nivel. Adicionalmente, el Cuadro **6-2** expone los resultados de la evaluación conceptual mediante la tarea de agrupamiento semántico; cada intervalo tiene un 95 % de confianza. Este estudio estableció una referencia manual de los documentos agrupados para medir el índice frand ajustado. Esta referencia tenía grupos de documentos significativamente superpuestos.

LDA construyó conceptos más coherentes que otras particiones dentro de las experimentadas, sin

[4] `https://github.com/nicolasdugue/DirectedLouvain`

Cuadro **6-2**: Resultados agrupamiento semántico

Algoritmo	**Índice Dunn**	**Medida de superposición**	**Índice frand ajustado**
LDA	0.177 (0.00128)	78.287 (0.00466)	0.83 (1,686e−5)
Directed Louvain 1	0.024 (0.1707)	188.0662 (0.1161)	1.0 (0.00053)
Directed Louvain 2	0.1833 (0.1273)	160.6315 (1.322)	0.9863 (0.00245)
Directed Louvain 3	0.309 (0.0901)	140.3709 (0.8845)	0.9201 (0.00036)

embargo, este resultado debe analizarse con cuidado ya que una análisis manual de los conceptos permite determinar que las unidades no describen constructos diferenciables entre sí. El Cuadro **6-3** expone los 15 top términos de los conceptos LDA, en particular, se señala con colores los términos repetidos dentro de las unidades conceptuales.

Cuadro **6-3**: Composición de conceptos generados con algoritmo LDA

15 top términos de conceptos LDA		
mujer	negocio_inclusivo_cacao	delicia_sexual
trabajar	constructor	alterno
numeroso	sonar	paz
cumbre_regional_paz	vehiculo	principio
victima	movil	conflicto_armar
preguntar	cosecha	seccional
unidad_reaccion_inmediato	tristeza	representacion
adulto	soledad	clase
impulsar	monte_marianas	impulsor
territorio_nacional	armar	pared
mujer	vida	comunicar
ejecutivo	concluir	fortalecer
victima	acceso_justicia	pacto_municipal
turismo	farc	colombia
protesta	conflicto_armar	jugar
mujer	poder	nino
estar	violencia_sexual	arauca
victima	pais	conflicto_armar
colombia	paz	caso
ano	farc	deber
farc	sufrir	colombia
memoria_dolor	valle_cauca	constructor
conflicto_armar	resguardo_indigena	uniforme_militar
torturar	vinculo	paz
tasa	relevante	desaparicion_involuntario

*Los términos presentados surgen del preprocesamiento descrito en el capítulo 5.2 donde se efectuó lematización y eliminación de signos diacríticos.

Es necesario señalar que los términos repetidos en los conceptos son palabras que aparecen en

la mayoría de los textos, por ejemplo, *conflicto_armar* es un término principal en cuatro de los cincos conceptos y está presente en 139 de los textos, así mismo, *mujer*, *víctima* y *paz* aparecen en tres unidades conceptuales cuando estas palabras respectivamente surgen en 194, 148 y 139 de los documentos.

En este sentido, las estructuras conceptuales poseen términos comunes que co-ocurren frecuentemente dentro de la mayoría de documentos, produciendo un alto nivel de coherencia, aun cuando los conceptos no facilitan segregar los documentos en conglomerados densos (ver Cuadro **6-2**). Esta situación permite establecer dos inferencias. La primera es que TC es una métrica sesgada cuando todos los conceptos están conformados por las mismas palabras frecuentes, por ende, no es apropiado utilizar esta medida como únicamente fuente para valorar el desempeño del reconocimiento de temáticas. La segunda deducción es que los conceptos LDA aquí generados, no son constructos adecuados para representar la información del corpus analizado.

Los resultados de LDA pueden ser consecuencia del tamaño del corpus dado que investigaciones como [Syed and Spruit, 2017] señalan que este algoritmo tiende a generar conceptos *gross-grained* en corpus pequeños, en particular, los autores reportan que LDA detalla los conceptos en corpus que tiene más de 142,000 tokens, mientras que el corpus aquí trabajado cuenta con 103,635 tokens.

El Cuadro **6-4** plantea algunas características numéricas de los conceptos producidos con el algoritmo *Directed Louvain* con intervalos de confianza al 95 %. Igualmente, el Cuadro **6-5**, expone 10 top términos de algunos conceptos *Directed Louvain* 1, mientras que el Cuadro **6-6** presenta esta información para *Directed Louvain* 2 y el Cuadro **6-7** lo hace para *Directed Louavin* en el tercer nivel jerárquico.

Cuadro **6-4**: Características numéricas de conceptos generados con algoritmo *Directed Louvain*

	Directed Louvain 3	**Directed Louvain 2**	**Directed Louvain 1**
No. conceptos	18.56 (0.4213)	55.4 (1.91)	941 (10.223)
No. términos promedio	334.362 (7.3392)	112.9358 (4.131)	6.571 (0.069)

Los conceptos de *Directed Louvain* en su primer nivel jerárquico tienen el menor nivel de coherencia, es decir, los términos que forman cada unidad conceptual no tienden a ocurrir en los mismo documentos. Esto era previsible porque son conceptos conformados por pocos términos en comparación con las otras particiones (ver Cuadro **6-4**), por ende, tienen menor probabilidad de mejorar su TC a medida que aumenta la ventana de los top términos examinados. Además, el Cuadro **6-5** permite evidenciar que los conceptos no describen estructuras claras y diferénciales entre sí, produciendo que los documentos tengan un nivel de asociación muy similar a cada unidad conceptual y el mayor valor de superposición durante el agrupamiento (ver Cuadro **6-2**). Esto advierte que este escenario posee conceptos cuya información semántica no es diferenciable y causa que los textos no puedan ser segregados. Por lo anterior, *Directed Louvain* 1 no produce conceptos y relaciones con los cuales describir el corpus.

El Cuadro **6-6** permite evidenciar que algunas estructuras conceptuales *Directed Louvain* 2 son conjuntos de términos con características comunes entre sí y diferentes a otras conglomeraciones, por ejemplo, el concepto tres contiene palabras que podrían describir aspectos productivos[5]. No obstante, la coherencia de los conceptos es baja, en este sentido, los conceptos construidos no explican constructos diferenciables entre sí. Lo anterior produce que los documentos tengan una

[5]Esta interpretación surge de analizar visualmente los conceptos y no implica un juicio de experto, por tal razón, la afirmación contiene la palabra "podría".

Cuadro **6-5**: Composición conceptos *Directed Louvain* 1

10 top términos de conceptos Directed Louvain 1	
violencia_sexual	marco_conflicto
mujer_victima	abuso_sexual
marco_conflicto_armar	reclutamiento_forzar
acceso_carnal_violento	alto_comisionado_paz_nacion
instituto_nacional_medicina_legal	delito_integridad_sexual
ley_victima_restitucion_tierra	desestigmatizacion
restitucion_tierra	alternativa_viable
reparacion_colectivo	ley_amnistia
bernardo_cuero_bravo	aprobar
enfoque_transformador	falencias
conflicto_armar	coalicion_nino_joven
mujer_desplazar	rol_mujer
cuerpo_mujer	estar_colombia
mujer_sobreviviente	vivir_carne
victima_violencia	region_castigar
organizacion_nacion_unir	victima_conflicto_armar
resgistro_unico_victima	violencia_genero
mesa_mujer_paz_seguridad	corporacion_sisma_mujer
contexto_conflicto_armar	vida_libre_violencia
ruta_pacifico_mujer	victimizar
proceso_paz	guerilla_eln
implementacion_acuerdo_paz	farc
gobierno_nacional	dejacion_arma
frente_domingo_lain_saenz	grupo_guerrillero
gobierno_colombia	disidencia_farc
defensor_derecho_humano	socorro_ramirez
lideresas_social	tratar_mujer
excombatiente_farc	persona_asesinar
lideresa	sangre_frio
dirigente_comunitario	lider

*Los términos presentados surgen del preprocesamiento descrito en el capítulo 5.2 donde se efectuó lematización y eliminación de signos diacríticos.

posibilidad semejante de pertenecen a cualquier partición (ver Cuadro **6-2**). En este orden de ideas, las estructuras conceptuales de *Directed Louvain* en el segundo nivel jerárquico no son particiones coherentes y no facilitan la agrupación de textos.

Por otro lado, los conceptos resultantes con *Directed Louvain* 3 poseen la mayor coherencia, además permiten un agrupamiento denso y muy similar a la referencia manual. Así, este escenario está formado por conceptos que describen constructos diferenciables entre sí, por ejemplo, el Cuadro **6-7** exhibe el concepto tres donde se asocian términos que pueden describir lugares geográficos de Colombia, mientras que el concepto cinco podría apuntar a los actores del conflicto armado.

Los hallazgos sobre *Directed Louvain* son afines a los reportados en [Li et al., 2013] donde agrupan semánticamente los documentos mediante un algoritmo de optimización de modularidad, produciendo grupos de textos con baja superposición y muy similares a la referencia manual. Igualmente, en [Ping and Chen, 2018] los autores emplearon un algoritmo Louvain para la construcción de conceptos para la narración visual de artículos científicos.

Cuadro **6-6**: Composición conceptos *Directed Louvain* 2

10 top términos de conceptos Directed Louvain 2	
conflicto_armar	psicologico
conflicto	impacto
esfuerzo	expresar
importante	fisico
asegurar	causa
muerte	camarada
causar	aniversario
dano	yira_medina
reacomodo	ofrendar
individuo	sandra_velez
artesania	elaboracion
mortalidad_materno	piscicultura
arbol	aguacate_hass
encadenar	reporteria
porcicultura	desastre_natural
problema	subregistro
solucion	corrupcion
resolver	distincion
materializar	garantia_independencia
recolectar_dato	ambito_nacional
vocero	program_desarollo_rural
plan_nacional_accion	programa_prevencion
marcha	patriotico
secundar_hombre	puesta
apoyo_proceso_negociacion	funcionamiento
hora	agregar
campana	vuelta
voto	salud_sexual_reproductivo
presidencia	batalla
presidencial	independencia_economico

*Los términos presentados surgen del preprocesamiento descrito en el capítulo 5.2 donde se efectuó lematización y eliminación de signos diacríticos.

No obstante, los resultados presentados son diferentes a los reportados en [Jia et al., 2018, Chen et al., 2012]. Estas investigaciones utilizaron algoritmos de agrupación particional (siguiendo la clasificación de [Fortunato, 2010]) para formar comunidades, en este sentido, generaron conglomerados de documentos densos y muy parecidos a la referencia manual. Las agrupaciones particionales se caracterizan por llevar los términos (del vocabulario) a un espacio métrico donde se determina una medida de distancia para formar comunidades. Los estudios nombrados pueden ser cuestionados ya que utilizan la frecuencia de los términos en el corpus para la formación de conceptos, desestimando las relaciones semánticas que pueden existir entre las palabras [Ali and Melton, 2018], por el contrario, el trabajo aquí documentado utilizó los arcos de la red de palabras para capturar información semántica. Adicionalmente, el empleo de agrupaciones particionales no es apropiado porque puede ser sensible a modificaciones en la medida de distancia usada durante la construcción de comunidades [Fortunato, 2010].

Este trabajo empleó una metodología, para agrupar semánticamente los documentos, similar a la

Cuadro **6-7**: Composición conceptos *Directed Louvain* 3

10 top términos de conceptos Directed Louvain 3	
victima	reconocer
violencia_sexual	cometer
violencia	violencia_mujer
forma	crimen
ley	mujer_victima
economico	problema
social	seguridad
cultural	condicion
politico	capacidad
implicar	accion
arauca	municipio
antioquia	choco
departamento	ciudad
bogota	narino
cauca	meta
nacional	local
coordinador	joel_sierra
derecho_humano	contraloria_general_republica
vocero	congreso
internacional	derecho_internacional_humanitario
farc	habano
eln	grupo_armar
gobierno	acuerdo_paz_colombia
guerrilla	acuerdo
paramilitar	punto
informe	jurisdiccion_especial_paz
defensoria_pueblo	colectivo
cnmh	documento
explicar	representante
fiscalia_general_nacion	especial

*Los términos presentados surgen del preprocesamiento descrito en el capítulo 5.2 donde se efectuó lematización y eliminación de signos diacríticos.

presentada en [Liu et al., 2013]. Sin embargo, el estudio aquí expuesto clasifica los términos (nodos) dentro de cada concepto (comunidad) usando la métrica WLR. Por lo anterior, se considera que la relevancia de un nodo está determinada por el número y el peso de sus vecinos [Lü et al., 2011]. Por el contrario, en [Liu et al., 2013] sólo se cuantifica el grado del nodo, generando que un término relevante sea aquel enlazado a muchos nodos que pueden no ser influyentes.

Las diferencias presentadas pueden ser la razón por la cual en [Liu et al., 2013] los autores establecieron que los conceptos extraídos mediante *clique percolation method* eran los apropiados para la agrupación semántica de textos. Igualmente, en [Liu et al., 2013] evaluaron la calidad de los conceptos construidos contra un listado de referencia manual. En este trabajo no se realizó esta validación debido a que, como ya se ha mencionado, este recurso no existe para el dominio estudiado y su construcción implicaba altos costos que justamente se deseaban eludir con el propósito de proponer una evaluación económica y aplicable a dominios que no poseen bases conceptuales.

El Cuadro **6-8** resume la propuesta para construir y evaluar estructuras ontológicas, comparándola

con investigaciones similares que han sido tratadas en la sección 6.3. En particular, el Cuadro **6-8** caracteriza el tipo de red que analizan las investigaciones, las estrategias para extraer y validar estructuras ontológicas, así como un resumen de los resultados reportados. El Cuadro **6-8** brinda información descriptiva ya que los estudios varían respecto al tipo de evaluación y métricas utilizadas para valorar el desempeño de las estrategias.

6.4. Depuración de conceptos

Una falencia del resultado *Directed Louvain* 3 es el alto número de términos dentro en los conceptos (ver Cuadro **6-4**), en este sentido, no es factible que un ser humano analice todos los términos asociados a cada unidad conceptual para brindar una interpretación. En este sentido, los conceptos fueron depurados siguiendo procedimiento descrito en la sección 5.3.3 donde el proceso se detiene cuando la coherencia del concepto en la iteración $i + 1$ menos la TC del concepto en la iteración previa i alcanza un valor d; dicho parámetro es establecido por un usuario externo.

Este trabajo experimentó con valores de 0.5, 1.0 y 1.5 para el parámetro d. El Cuadro **6-9** expone el número de términos resultantes en cada escenario y el número de palabras promedio en cada conceptos, consideran intervalos de confianza al 95 %.

El Cuadro **6-9** permite señalar que a menor valor de d, menor número de términos tendrá cada concepto, por lo anterior, se debe establecer un valor pequeño para este parámetro si un humano deseará interpretar los conceptos generados. Igualmente, la Figura **6-2** expone la coherencia de los conceptos *Directed Louvain* 3 originales y depurados, con una barra de error al 95 % de confianza.

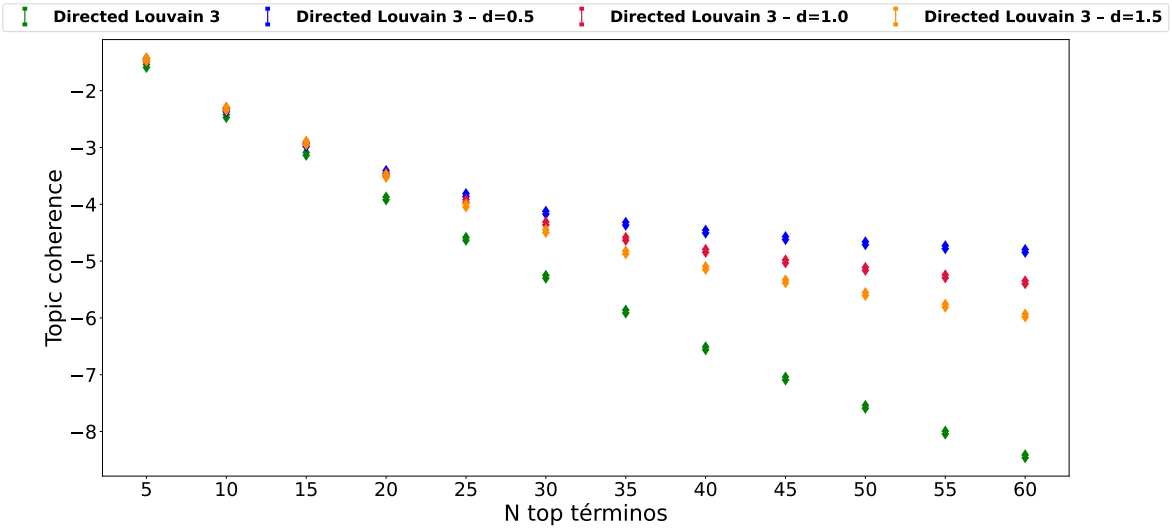

Figura **6-2**: Coherencia del Tópico (*Topic coherence* - TC) con top términos entre 5 a 60 en conceptos *Directed Louvain* 3 originales y depurados.

La Figura **6-2** permite señalar que la depuración propuesta, como se esperaba, afecta a los términos menos relevantes dentro de los conceptos, por esta razón, no existe una diferencia significativa en la coherencia entre los diferentes escenarios para los 30 primeros top términos.

Por lo anterior, ejecutar la prueba *word intrusion* con los diez primeros términos de cualquiera de

Cuadro **6-8**: Comparación entre la propuesta planteada y estudios similares.

Estudio	**Tipo de red**	**Extracción estructuras ontológicas**	**Validación estructuras ontológicas**	**Resumen resultados**
[Chen et al., 2012]	No dirigida	Algoritmo de agrupación particional	*Gold standard*	Recuperación con alto *recall* Estructuras precisas
[Li et al., 2013]	No dirigida	Algoritmo de optimización de la modularidad	*Gold standard*	Agrupación de textos con baja superposición Estructuras similares a la referencia manual
[Liu et al., 2013]	No dirigida	*Clique percolation method*	Evaluación mediante agrupamiento semántico (métrica grado del nodo) *Gold standard*	Estructuras precisas Textos agrupados en particiones densas con baja superposición
[Jia et al., 2018]	Dirigida	Algoritmo de agrupación particional	*Normalized Pointwise Mutual Information* *Gold standard*	Estructuras precisas Asociación entre entidades
[Ping and Chen, 2018]	No dirigida	Algoritmo de optimización de la modularidad	Evaluación basada en humanos	Según el criterio de los encuestados, estructuras describían entidades y asociaciones
[Qiu et al., 2020]	No dirigida	Algoritmo de optimización de la modularidad	*Gold standard* Modularidad	Estructuras con alta modularidad y F-measure
Propuesta	Dirigida	Algoritmo de optimización de la modularidad	*Topic coherence* Evaluación mediante agrupamiento semántico (métrica WLR)	Conceptos coherentes (ver Figura **6-1**) Textos agrupados en particiones densas con alto índice frand ajustado (ver Cuadro **6-2**)

Cuadro **6-9**: Características numéricas de conceptos *Directed Louvain* 3 depurados

	d=0.5	***d=1.0***	**d=1.5**
No. términos	767.6 (19,90547)	936.666 (23.9355)	1191.1 (28.86147)
No. términos promedio	41.464 (1.1171)	50.6506 (1.55066)	64.5141 (2.2226)

estos escenarios, brinda una idea de cómo los seres humanos perciben la precisión de las particiones. Este trabajo utilizó como evaluadores a tres estudiantes de ingeniería. El Cuadro **6-10** expone la precisión que los sujetos encuestados adjudicaron a cada concepto, en este sentido, C4[6] y C5[7] estaban formados por términos que permiten describir un constructo particular, por otro lado, C14 es una estructura imprecisa pues sus palabras relevantes eran *(desplazar, asesinar, persona, octubre, periodo)* que pueden indicar diferentes conceptos.

Sin embargo, la evidencia planteada en el Cuadro **6-10** debe analizarse con cuidado dado que la prueba implementada, y documentada en [Chang et al., 2009], no permite a los encuestados cuantificar el grado de interpretabilidad de los conceptos, por el contrario, la decisión es binaria, así penaliza a constructos donde la mayoría de sus términos son coherentes y sólo unos pocos son inadecuados [Senel et al., 2018].

Cuadro **6-10**: Resultados de *word intrusion* para una partición *Directed Louvain* 3

Concepto	**Precisión**	**Concepto**	**Precisión**
C1	0.556	**C8**	0.222
C2	0.333	**C9**	0.333
C3	0.778	**C10**	0.333
C4	1.0	**C11**	0.444
C5	1.0	**C12**	0.222
C6	0.222	**C13**	0.222
C7	0.444	**C14**	0.0

La propuesta implementada para depurar conceptos es diferente a la planteada en [Dang and Nguyen, 2018] que eliminan los términos menos relevantes según la medida *betweenness*, por ende, este trabajo desestima la influencia que el nodo recibe de su vecindario [Lü et al., 2016], es así como el estudio es criticable porque ignora el contexto donde se encuentran las palabras.

Utilizar información semántica permite organizar los términos considerando relaciones que los datos tal vez no describen, por ejemplo, en [Eissa et al., 2018] descartan los nodos que no comparten características semánticas de la mayoría de vértices dentro de las comunidades, en especifico, los autores perfilan personas mediante redes de palabras y los atributos semánticos que usan es el lugar de nacimiento, formación académica, etc.

La investigación aquí propuesta no emplea características semánticas porque la única fuente estructurada del dominio (el tesauro del CNMH) no contiene atributos semánticos para la mayor parte del vocabulario analizado. Una opción sería filtrar las estructuras conceptuales dependiendo de la etiqueta POS de sus términos, no obstante, este criterio supondría que los conceptos son descritos exclusivamente por verbos o sustantivos, cuando esta implicación no es cierta ya que el tesauro del CNMH tiene definiciones como la exhibida en el párrafo (6.4-a) donde los sustantivos se señalan en color naranja y los verbos en azul. El etiquetado del párrafo (6.4-a) surge de procesar el texto con

[6]Este concepto contiene top términos como *(arauca, antioquia, departamento, bogota, cauca)*.
[7]Los términos más relevantes, de este concepto, son *(farc, eln, gobierno, guerrilla, paramilitar)*.

el *parser* de Freeling 4.0.

(6.4-a)"**Arma trampa:** Todo artefacto o material concebido construido o adaptado para matar o herir y que funcione inesperadamente cuando una persona toque un objeto aparentemente inofensivo o se aproxime a él, o realice un acto que aparentemente no entrañe riesgo alguno."[Espinosa, 2018, p. 18].

Otra opción para depurar los conceptos mediante atributos lingüísticos es expuesta en [Bunk and Krestel, 2018] donde se propone enriquecer semánticamente los términos al aprender la distribución del vocabulario en el modelo *word2vec* pre-entrenado con los documentos de *Google News*, así los autores ubican en mejores posiciones a las palabras que sean más similares a los términos top siguiendo la información del modelo embebido, de esta forma, el estudio reporta valores superiores a 0.5 en la prueba *word intrusion*. En [Bunk and Krestel, 2018] implementan esta estrategia porque se manipuló el corpus *20News* que es una colección de noticias en inglés recuperadas a finales del siglo pasado, por ende, el modelo pre-entrenado contiene gran parte del vocabulario de interés y exhibe relaciones para incrementar las características semánticas de los términos. Por el contrario, la propuesta planteada en el capítulo 5 no aplicó el trabajo expuesto pues los actuales modelos embedidos en español carecen de la terminología especializada del conflicto armado colombiano.

La propuesta documentada requiere que un usuario establezca el parámetro d para frenar la depuración de conceptos. La intervención humana es necesaria porque hasta el momento no se ha encontrado en la literatura alguna evaluación basada en datos que permita definir cuándo detener este proceso. Así mismo, esta estrategia ha sido empleada en investigaciones semejantes y nombradas previamente [Dang and Nguyen, 2018, Eissa et al., 2018]. Inicialmente, la autora intento utilizar el agrupamiento semántico para valorar el desempeño de los conceptos depurados bajo diferentes rangos de d.

El Cuadro **6-11** presenta los resultados de la agrupación de textos considerando tres valores del parámetro d y variando el número de términos top empleados durante la tarea. Esta información permite señalar tres aspectos. El primero es que la depuración propuesta genera conceptos donde los documentos pueden ser agrupados, más aún dichos conglomerados son más densos en comparación con las particiones *Directed Louvain* 3 originales. No obstante, el segundo aspecto es que no existe diferencia estadísticamente significativa al modificar el valor de d, por ende, la tarea no es una evaluación apropiada para calibrar este parámetro. El tercer aspecto apunta a que la tarea de agrupamiento semántico beneficia a las particiones con un mayor número de términos, por ejemplo, el índice Dunn siempre es mejor para 50 top términos que en el escenario de 30 palabras.

El último aspecto ocurre porque la asociación entre conceptos y documentos se cuantifica a través de una similitud coseno, por lo anterior, entre más términos posean las entidades (tanto los textos como las unidades conceptuales) existirá una mayor afinidad entre estas. Lo anterior no implica que la evaluación propuesta por este trabajo sea inadecuada para validar estructuras ontológicas porque, como se expuso en párrafos previos y se nota en el Cuadro **6-2**, el agrupamiento semántico es capaz de valorar el desempeño de particiones donde el número de unidades conceptuales y la composición de los conceptos (*i.e.* términos agrupados) son diferentes. Sin embargo, la tarea propuesta no es sensible cuando las estructuras conceptuales son muy similares como las tres particiones *Directed Louvain* 3 depuradas.

Cuadro **6-11**: Resultados agrupamiento semántico para conceptos *Directed Louvain* 3 depurados

	d= 0.5		
No. top términos	**Índice Dunn**	**Medida de superposición**	**Índice frand ajustado**
10	0.3045 (0.03964)	148.8176 (1.0722)	0.9047 (0.00014)
20	0.3459 (0.04266)	141.2176 (1.0388)	0.9145 (0.000176)
30	0.587 (0.04451)	136.0442 (1.17306)	0.9189 (0.000167)
40	0.6975 (0.0505)	131.7228 (1.1934)	0.92034 (0.000189)
50	0.7012 (0.0561)	127.8287 (1.20177)	0.9204 (0.000194)
	d= 1.0		
No. top términos	**índice Dunn**	**Medida de superposición**	**Índice frand ajustado**
10	0.3178 (0.03989)	148.6826 (0.9766)	0.89997 (0.00014)
20	0.351 (0.04534)	141.732 (0.90018)	0.9035 (0.00016)
30	0.58848 (0.03983)	137.3302 (0.92353)	0.9147 (0.000164)
40	0.70145 (0.04206)	133.45873 (1.05743)	0.91745 (0.000178)
50	0.70478 (0.05023)	130.2931 (1.099)	0.9247 (0.00019)
	d=1.5		
No. top términos	**Índice Dunn**	**Medida de superposición**	**Índice frand ajustado**
10	0.320143 (0.038)	149.1305 (0.9307)	0.90991 (0.000133)
20	0.3719 (0.04925)	142.3553 (0.95311)	0.91873 (0.000142)
30	0.60337 (0.04058)	138.2522 (0.97665)	0.92418 (0.000149)
40	0.71447 (0.04083)	135.3074 (1.02373)	0.92412 (0.000157)
50	0.71543 (0.01513)	132.1916 (0.33694)	0.9347 (0.000045)

Capítulo 7

Conclusiones

Este trabajo expone una metodología de enriquecimiento ontológico enfocada en el aprendizaje ontológico a partir de textos en español que abordaban el dominio del conflicto armado colombiano, así es como este estudio es el primero en presentar una propuesta computacional para la extracción y evaluación semi automática de conceptos y relaciones para este dominio. Los hallazgos expuestos pueden ser la base para la refinar una ontología en este campo. Es necesario divulgar que esta investigación cumplió en totalidad los objetivos propuestos, a continuación, se plantean las conclusiones encontradas para cada uno de los objetivo planteados.

1. Preparar los datos textuales que describen las graves violaciones a los derechos humanos e infracciones al derecho humanitario con ocasión del conflicto armado colombiano.

Esta investigación recuperó y procesó un conjunto de noticias en español que aborda el conflicto armado colombiano; igualmente, formalizó el tesauro del CNMH. Estos recursos pueden utilizarse fácilmente en futuras investigaciones enfocadas en construir un lenguaje común que caracterice este dominio. Así, las instituciones estatales podrían describir comunidades de interés sin necesidad de invertir una gran cantidad de recursos (*i.e.* tiempo, humanos, dinero, etc.). En las secciones 5.1 y 6.1 están documentados los procesos que sustentan la ejecución del primer objetivo.

2. Formular algoritmos para el aprendizaje ontológico considerando las limitaciones de recursos lingüísticos que tiene el idioma español y el dominio del conflicto armado colombiano.

Dadas las condiciones experimentales, el vocabulario construido al modificar el estudio de Meijer et al. [2014] posee el valor *F-measure* más alto ya que recupera un mayor número de colocaciones y unigramas que realmente pertenecen al *gold standard*. Los conceptos detectados mediante el algoritmo *Directed Louvain* en el tercer nivel jerárquico generaron conglomerados densos con un índice de superposición de $140{,}3709 \pm 0{,}8845$, siendo estos conglomerados muy parecidos a la referencia manual (ver Cuadro **6-2**). Además, el resultado TC indica que los conceptos extraídos son coherentes. Por ende, el algoritmo *Directed Louvain* 3 detecta conceptos y relaciones apropiadas para describir el corpus en español que aborda el dominio del conflicto armado colombiano (ver Figura **6-1**).

Igualmente, esta investigación depura los conceptos mediante un proceso que busca maximizar la coherencia de las estructuras conceptuales. En particular, la depuración de conceptos elimina muchos de los términos del vocabulario (el 80 % cuando *d=1.5* ver Cuadro **6-9**), es decir, la mayor parte de las palabras extraídas no son utilizados para formar conceptos. Por lo anterior conceptos

finos (*fine-grained*), que pueden existir dentro de los términos eliminados, no son recuperados. Por ejemplo, ninguna de las particiones resultantes *Directed Louvain 3 -depuradas* contiene los términos *(etnoturismo, agroturismo, ecoturismo, turismo comunitario)* que podrían describir el constructo *zonas de interés de desarrollo rural, económico y social*[1], aún cuando estos términos si existen en los conceptos *Directed Louvain* 3 originales.

Lo anterior implica que la propuesta documentada facilita construir estructuras ontológicas diferenciables entre sí y que favorecen la agrupación de documentos, no obstante, los conceptos resultantes capturan las temáticas predominantes dentro del corpus y no constructos específicos. Este aspecto es una falencia de la actual propuesta ya que el objetivo es recuperar entidades conceptuales que permitan enriquecer una ontología, por consiguiente, omitir conceptos detallados conlleva que la esquematización no será refinada con toda la información que está contenida en los textos.

3. Evaluar las estructuras ontológicas que sean producto de aplicar los algoritmos establecidos.

La evaluación con *gold standard* para validar el vocabulario fue ventajosa y asequible porque los costos de construir la referencia manual sólo ocurrieron una vez [Dellschaft and Staab, 2008]. Sin embargo, este enfoque conlleva que la referencia terminológica describa únicamente los documentos analizados. Por ende, el *gold standard* es un insumo que no responde a la naturaleza dinámica de la información [Flouris et al., 2008] ya que no contiene datos que puedan surgir en la posteridad [Clark et al., 2012]. Igualmente, los resultados de la evaluación con *gold standard* podrían verse afectados por las métricas de *precision* y *recall* empleadas ya que cuantifican el emparejamiento perfecto entre los términos extraídos y los presentes en la referencia, es decir, no consideran variaciones de los términos como la sinonimia.

Este trabajo plantea la evaluación de las estructuras ontológicas abstractas mediante la tarea de agrupación semántica de documentos. La propuesta de evaluación, a diferencia de las ya existentes, permite validar conceptos y relaciones, sin la costosa necesidad de utilizar evaluación por humanos ni bases conceptuales. Por ende, esta propuesta no sólo es económica sino también aplicable para modelar datos y dominios que carecen de fuentes de conocimiento estructurado.

Esta investigación propone cuantificar automáticamente los resultados de la agrupación de textos a través de métricas de densidad, superposición y similitud con datos de referencia, además, medir automáticamente la coherencia de los conceptos identificados. Así, este trabajo se compromete con la evaluación a bajo costo y valoración a gran escala. El enfoque propuesto no es dependiente de información del dominio y satisface los criterios indicados por [Dellschaft and Staab, 2008] que permiten filtrar la intervención de la tarea para obtener conclusiones válidas sobre las técnicas de aprendizaje ontológico que se comparan en este trabajo. No obstante, la evaluación propuesta no es sensible para valorar particiones con una composición (*i.e.* orden de los términos dentro de los conceptos) muy semejante, es así como esta tarea no es apropiada durante la depuración de conceptos (ver sección 6.4).

Las métricas, presentadas en la sección 5.3.2, permitieron caracterizar robustamente las técnicas para extraer estructuras ontológicas abstractas pues se cuantificó la coherencia de los conceptos, además la densidad, superposición y similitud de la agrupación semántica de los textos. Así se identificó que LDA genera tópicos con los cuales se producen grupos de textos con bajos niveles de superposición a diferencia de los algoritmos Louvain analizados, no obstante, los conceptos construidos con LDA aún cuando tienen un alto TC no son diferenciables entre sí ya que las unidades conceptuales comparten términos top.

[1] El concepto señalado está descrito en el tesauro del CNMH.

Capítulo 8

Futuros trabajos

El trabajo abordado puede ser extendido y profundizado de diferentes formas. Este apartado plantea guías para futuras investigación considerando los datos analizados, así como la extracción y evaluación de estructuras ontológicas. Sumado a esto, se presenta una breve sección donde se discuten trabajos que conjugan enriquecimiento ontológico y *word embedding*, así se brindan guías sobre cómo futuras investigaciones pueden aprovechar estas estrategias que han mejorado la comprensión de textos y la ingeniería de conocimiento [Al-Aswadi et al., 2020].

1. Datos analizados

El corpus construido y el tesauro formalizado del CNMH pueden ser la base para ejecutar trabajos orientados en las dos últimas etapas del enriquecimiento ontológico, es decir, en la sugerencia y validación de cambios ontológicos.

2. Extracción de estructuras ontológicas

En trabajos futuros, escenarios experimentales con otros esquemas de ponderación locales y globales (ver sección 5.2) pueden ser considerados para analizar la influencia que la medida local tuvo en los resultados del Cuadro **6-1**. Así mismo, investigaciones futuras pueden utilizar patrones léxicos, si estos se construyen a través de herramientas no supervisadas o aplicando un enfoque semi-supervisado.

Siguiendo lo reportado en [Qiu et al., 2020], se propone explorar diferentes estrategias para construir la red que sirve como entrada para la detección de comunidades. Algunos parámetros que se pueden modificar son:

- Tamaño de la ventana de palabras: En [Levy and Goldberg, 2014] reportan que una ventana de cinco palabras permite construir unidades que comparten un contexto temático, mientras que una ventana de dos palabras (es decir, la empleada en la sección 6.3) captura elementos con una misma característica sintáctica (como la etiqueta POS). Por lo anterior, futuros trabajos podrían establecer escenarios experimentales variando este parámetro, de esta forma, evaluar el impacto en la coherencia de las estructuras conceptuales.
- Ponderación dentro de la ventana de palabras: La autora construyó la red sin ponderar los términos, es decir, al analizar el fragmento $[mujer\_victima, farc, disponer]$ se considero que

las tuplas $(mujer\_victima, farc)$ y $(farc, disponer)$ aportaban el mismo valor a la ocurrencia de estas unidades, aún cuando el primer conjunto está relacionado temáticamente [1].

Por ende, otros estudios podrían asignar una ponderación a tuplas que coocurran y compartan características semánticas, no obstante, este lineamiento implica explotar conocimiento de fuentes estructuradas, limitando la investigación a dominios previamente analizados.

Otra alternativa es adjudicar la ponderación considerando la proximidad de cada término con la palabra central dentro de la ventana o las relaciones sintácticas que comparten [Levy et al., 2015]. Estas modificaciones permiten que las estructuras conceptuales obtengan un alto desempeño ([Levy et al., 2015] reporta 60 % de precisión) durante tareas de analogías.

- Tipo de arco: Establecer arcos no dirigidos, es decir, asumir el modelo *bag-of-words* para diseñar la red permite que la matriz de coocurrencia sea menos dispersa. Futuros trabajos pueden cuantificar el impacto de utilizar esta estructura en la coherencia de los conceptos.

La propuesta documentada facilita extraer estructuras ontológicas que favorecen el agrupamiento de textos, sin embargo, los conceptos resultantes capturan las temáticas predominantes dentro del corpus. Esta falencia podría abordarse en futuras investigaciones que planteen la construcción de conceptos como una mezcla de las particiones generadas por modelos jerárquicos, como el algoritmo *Directed Louvain* aquí utilizado. Igualmente, es apropiado que futuros estudios examinen algoritmos de detección de comunidades considerando superposición, de esta forma, explotar la polisemia de los términos al permitir que una palabra conforme más de una unidad conceptual.

3. Evaluación de estructuras ontológicas

Otros estudios pueden considerar dos alternativas para aminorar los sesgos que tiene la evaluación con *gold standard* durante la validación del vocabulario. La primera es plantear tareas más intensivas como la identificación de sinónimos, la revisión ortográfica y la desambiguación lingüística durante el preprocesamiento de datos con el propósito de mejorar la calidad de los términos, por lo tanto, su rendimiento durante la evaluación. La segunda alternativa es utilizar medidas más robustas como las propuesta en [Dellschaft and Staab, 2008] donde se consideran variaciones léxicas, sin embargo, emprender este camino es una decisión cautelosa porque los hallazgos pueden ser poco comparables con otras investigaciones dada la popularidad de emplear las métricas de los sistemas de recuperación de información para evaluar la formación de vocabulario.

Futuras investigaciones pueden implementar las métricas basadas en redes que plantea [Dey et al., 2018] para contrastarlas con la medida TC, así robustecer la información de coherencia evaluada partir del corpus de trabajo. Lo anterior podría contra restar el sesgo que tiene TC hacía particiones que contienen términos top repetidos que co-occurren frecuentemente en los textos. Trabajos que conozcan los términos relevantes dentro de los conceptos, podrían aplicar la métrica de [Senel et al., 2018] para reemplazar la prueba *word intrusion* ya que esta propuesta no requiere evaluación por humanos, además, emplea el mismo marco teórico donde se supone que los humanos interpretan los conceptos tratando de agrupar las palabras más distintivas dentro de cada partición. Igualmente, estudios posteriores pueden aplicar la propuesta de enriquecimiento enfocada en aprendizaje ontológico para modelar las estructuras conceptuales de un corpus ampliamente analizado, de esta manera, contrastar el desempeño de técnicas usuales para esta tarea y el planteamiento aquí abordado.

[1] Se afirma la relación temática entre *mujer_victima* y *farc* porque estos términos aparecen dentro del tesauro del CNMH.

4. Enriquecimiento ontológico y *word embedding*

Los modelos *word embedding* plantean representaciones vectoriales densas en una dimensión inferior al tamaño del vocabulario, donde los vectores capturan propiedades sintácticas y semánticas de los términos [Lane et al., 2019]. Es así como estos tienen un mejor desempeño que las transformaciones heurísticas en la matriz de coocurrencia[2] durante tareas de consulta y razonamiento lógico [Baroni et al., 2014].

Lo anterior ocurre porque los pesos dentro del vector *word embedding* v_i asociado al término t_i, se establecen de tal forma que se maximice la probabilidad de predecir la ocurrencia de las palabras cercanas a t_i (*i.e.* dentro de una ventana de palabras cuyo tamaño es determinado previamente), de esta forma, el modelo asigna pesos similares a términos que ocurren en contextos similares. Estos modelos estiman los vectores de palabras a través de una tarea supervisada que procesa un corpus grande[3] sin etiquetas porque la información contextual se extrae automáticamente del corpus.

Existen investigaciones que explotan los modelos *word embedding* para el aprendizaje y alineamiento de estructuras ontológicas, dado que estos modelos construyen vectores que describen el significado de los términos sin costos de anotación manual. Dichas investigaciones pueden clasificarse en tres categorías. En la primera, los estudios plantean estrategias para generar vectores de palabras al analizar textos de dominios específicos. La segunda categoría agrupa propuestas sobre representaciones vectoriales partiendo de ontologías formales o grafos de conocimiento general (*i.e.* YAGO, DBpedia, etc.). En la segunda se asocia con *inductive transfer learning* ya que los trabajos utilizan vectores pre-entrenados para abordar tareas ontológicas. El Cuadro **8-1** expone algunas investigaciones de enriquecimiento ontológico y *word embedding* siguiendo la estructura planteada en la sección 3.3.

En la primera categoría hay investigaciones como [Jiang et al., 2020] donde los autores generan vectores de palabras al emplear el modelo *word2vec* y ajustar la función objetivo para incorporar conocimiento de la similitud que dos términos poseen en la ontología MeSH, de esta forma, garantizar que información de las relaciones ontológicas sea captura en las representaciones vectoriales y no dejar esto a manos de una posible coocurrencia o proximidad dentro del corpus. Un estudio semejante es [Albukhitan et al., 2017] quienes construyen vectores densos de términos en árabe, posteriormente los autores establecen conceptos y relaciones taxonómicas al cuantificar la similitud coseno entre palabras del corpus y una lista de entidades generada por humanos, además los autores entrenan un clasificador para establecer relaciones no taxonómicas. Igualmente en [Denaux and Gomez-Perez, 2019] abordan el problema de ambigüedad de algunos modelos *word embedding*, para esto los autores formaron vectores de palabras a partir de un corpus donde los términos tenían etiquetas relacionadas a conceptos formalizados en BabelNet.

Las investigaciones de la primera categoría modifican los algoritmos para establecer los vectores de palabras, así enriquecer las representaciones resultantes con información semántica de contextos específicos. No obstante, estas modificaciones son dependientes a conocimiento externo, por ende, estos estudios limitan su impacto a dominios previamente analizados y la disponibilidad de los datos. Por ejemplo, en [Jiang et al., 2020] refinaron la información de 4,878 tokens de los 145,274 que existían en el corpus, ya que la ontología MeSH sólo esquematizaba estas unidades léxicas. Situación similar ocurre en [Denaux and Gomez-Perez, 2019] donde los autores aplicaron su estrategia de desambiguación al 25 % de los términos dada la información contenida en BabelNet. Una alternativa es utilizar a expertos humanos [Albukhitan et al., 2017], sin embargo, los trabajos quedan restringidos al conocimiento de los peritos e incurren en altos costos para acceder a estos recursos.

[2] Algunas heurísticas que manipulan la matriz de coocurrencia son los esquemas *Pointwise Mutual Information* o *Positive Pointwise Mutual Information*, así como factorizaciones SVD o NMF [Levy et al., 2015].

[3] En [Mikolov et al., 2013] el tamaño final del corpus (después de filtrado) es de 692k tokens.

La segunda categoría agrupa investigaciones cuyo objetivo es llevar grafos de conocimiento a espacios vectoriales densos donde la información semántica sea descrita. En [Bel-Enguix et al., 2019] utilizan *node2vec* para transformar los nodos de una red generada por humanos a una representación vectorial. Este algoritmo construye el vector v_i del nodo n_i, de tal forma que los pesos maximicen la probabilidad de observar los nodos que pertenecen al vecindario de n_i. No obstante, este tipo de estudios son criticables porque los vectores resultantes codifican el conocimiento contenido explícitamente en el grafo, por ende, las representaciones aprenden la versión condensa y filtrada del fenómeno real. Por lo anterior, los vectores de palabras sólo proporcionan una fracción del conocimiento que se podría adquirir al modelar un corpus formado con textos extraídos de la web [Denaux and Gomez-Perez, 2019]. Esta situación puede ser contrarrestada si la red conceptual contiene todas las posibles características del fenómenos, es decir, al estudiar grafos robustos como DBpedia que caracteriza 4.58 millones de elementos a través de 38 millones de nodos.

Las investigaciones de la tercera categoría utilizan estructuras pre-entrenadas para representar los términos, al suponer que los vectores de palabras entrenados previamente en grandes corpus encapsulan la información contextual [Khadir et al., 2021]. Estos estudios se enfocan en tareas de aprendizaje y alineamiento ontológico. Por ejemplo, en [Gupta et al., 2017] agrupan los términos para formar conceptos cuya etiqueta es el synset común que compartan los elementos de cada conglomerado, además la extracción de relaciones taxonómicas ocurre mediante análisis de coocurrencia, así los autores no suponen que las asociaciones del corpus son capturadas por los vectores pre-entrenados. Por otro lado, en [Huang et al., 2020] diseñan una red neuronal para clasificar relaciones, siendo los datos iniciales los vectores pre-entrenados que representan los términos de la base de datos TACRED. En especifico, los autores codifican la relación entre el término t_i y t_j como la resta entre los vectores embedidos $v_i - v_j$ dada la matriz V de vectores pre-entrenados; así mismo, los autores modifican los pesos del vector resultante con información del corpus.

Utilizar los vectores pre-entrenados evita la necesidad de procesar textos para extraer información semántica. Es así como los autores de [Khadir et al., 2021] analizan 153,928 relaciones de WordNet para entrenar una red neuronal que clasifique asociaciones. La entrada del clasificador son vectores concatenados de las representaciones embedidas, por ejemplo para la tripleta (t_i, r, t_j) el vector inicial es $v_i \oplus v_j$. Bajo una línea similar, en [Nkisi-Orji et al., 2019] alienan entidades conceptuales al entrenar un clasificador de bosques aleatorios que tiene como entrada los conceptos representados a través de los vectores embedidos, es decir, los autores no procesan textos para asociar los elementos ontológicos.

Modelar un idioma o dominio a través de *word embedding* implica tomar decisiones en relación a los recursos disponibles, por ejemplo, si se desea analizar el dominio del conflicto armado colombiano no sería adecuado encaminar las investigaciones bajo la primera categoría pues conllevan utilizar fuentes de conocimiento estructurado de las cuales carece este dominio. Una alternativa es transformar el tesauro del CNMH a un espacio vectorial, de esta forma, usufructuar el conocimiento explícito de este dominio; sin embargo, los vectores resultantes deben utilizarse con cuidado porque la información explicita capturada sería limitada dado que el tamaño del tesauro (ver sección 6.1).

En este orden de ideas, investigaciones como las reportadas en la tercera categoría facilitan formar estructuras ontológicas a partir de textos en español que aborden el conflicto armado colombiano. Sin embargo, existen algunos puntos a considerar como la codificación de las relaciones mediante vectores pre-entrenados. Lo anterior es importante porque [Levy and Goldberg, 2014] reporto que la resta (*i.e.* $v_i - v_j$) y concatenación (*i.e.* $v_i \oplus v_j$) de vectores (es decir, los empleados en [Huang et al., 2020] y [Nkisi-Orji et al., 2019]) no son mecanismos que capturen la inferencia léxica entre los términos, por el contrario, estas codificaciones tienden a aprender que t_j posee rasgos de hiperonimia sin considerar su relación con t_i.

Cuadro **8-1**: Investigaciones de enriquecimiento ontológico relacionadas con *word embedding*.

Estudio	**Entrada**	**Elementos ontológicos**					**Descripción**	**Tipo de evaluación**	**Intervención usuario**
		Términos	**Conceptos**	**Relaciones taxonómicas**	**Relaciones no taxonómicas**	**Axiomas**			
[Jiang et al., 2020]	RadCore y MIMIC-II (145,274 tokens) Ontología MeSH	×	×				Construcción *word embedding* Similitud coseno	*Semantic relatedness*	Semi automática
[Albukhitan et al., 2017]	Textos propios en árabe (5,000 tokens) Conceptos y relaciones semilla	×	×	×	×		Construcción *word embedding* Similitud coseno Clasificación supervisada	*Gold standard*	Cooperativa
[Denaux and Gomez-Perez, 2019]	Wikipedia inglés de 2018 (824,000 tokens) BabelNet	×	×	×			Construcción *word embedding* Clasificación red neuronal	*Word similarity* *Word prediction* Evaluación basada en tarea	Semi automática
[Gupta et al., 2017]	Yahoo Finance (33,939 sentencias) WordNet	×	×	×			Vectores pre-entrenados Agrupamiento Análisis de coocurrencia Humanos expertos	*Gold standard* Evaluación basada en humanos	Cooperativa
[Huang et al., 2020]	SemEval 2010 - task 8 (10,717 tripletas) TACRED (106,264 tripletas)			×	×		Vectores pre-entrenados Clasificación red neuronal	*Gold standard*	Semi automática
[Khadir et al., 2021]	WordNet (153,928 tripletas)			×	×		Vectores pre-entrenados Clasificación red neuronal	*Gold standard*	Semi automática
[Nkisi-Orji et al., 2019]	Tesauro European Union multilingual (7,234 conceptos) Tesauro General Multilingual Environmental (5,220 conceptos)		×				Vectores pre-entrenados Clasificación supervisada	*Gold standard*	Semi automática
[Bel-Enguix et al., 2019]	Grafo construidos por humanos (28,656 nodos)		×				*Node2Vec*	*Semantic relatedness* *Word similarity* Evaluación basada en tarea	Cooperativa
[Li et al., 2020]	Wiki-KBP (151,091 instancias de las cuales 38,922 están etiquetadas) TACRED (106,264 tripletas)			×	×		Vectores pre-entrenados *Distant supervision* Algoritmo de ponderación de muestras basada en *meta-learning*	*Gold standard*	Semi automática

Otra consideración es que los vectores pre-entrenados pueden no capturar la información semántica del dominio, por lo anterior, aplicar una estrategia *fine-tunning* favorecería la extracción de estructuras ontológicas. No obstante, esta línea de investigación implica el acceso a muchos de textos del dominio con el fin de lograr un buen desempeño, por ejemplo, en [Howard and Ruder, 2018] reportan que el sistema de clasificación redujo su tasa de error en 10 puntos cuando se ingresaron más de 20,000 textos de dominio.

Igualmente, los estudios que conjugan enriquecimiento ontológico y *word embedding* extraen estructuras ontológicas a través de tareas de clasificación supervisada donde es necesario utilizar conjuntos de datos etiquetados (ver columna **Entrada** del Cuadro **8-1**). Esta circunstancia está más acentuada cuando se entrenan clasificadores basados en redes neuronales que han superado los resultados del estado del arte para muchas tareas [Al-Aswadi et al., 2020]. Por lo anterior, es un desafío aplicar estas estrategias en idiomas como el español y el conflicto armado colombiano.

El aprendizaje de múltiples instancias (en inglés *multiple instance learning*) puede ser una solución, siempre que se cuente con algunas etiquetas para los documentos. Por ejemplo, en [Wang et al., 2016] extraen oraciones para describir situaciones de protestas al establecer que cada documento era una bolsa etiquetada con el tipo de evento y las sentencias dentro del texto eran instancias de dicha etiqueta. Una falencia de este aprendizaje supervisado débil es la ambigüedad, es decir, instancias positivas de una misma clase pueden tener diferentes estructuras [Li et al., 2020]. Así mismo, el aprendizaje de múltiples instancias resulta ineficiente para extraer relaciones ya que esta tarea está enfocada en analizar cada sentencia, no el documento.

Futuras investigaciones podrían considerar propuestas como las de [Li et al., 2020] donde clasifican relaciones siguiendo el supuesto *distant supervision*[4] para automáticamente generar datos de entrenamiento mediante textos sin etiquetas y una base de datos de relaciones. Más aún futuros trabajos pueden conjugar los beneficios de *distant supervision* y algún algoritmo de *meta-learning* como la ponderación de muestras a partir de datos etiquetados, de esta forma, reducir la retención de instancias de entrenamiento ruidosas (es decir, aquellas sentencias que contienen las entidades pero no describen relaciones de interés), así mejorar la capacidad del clasificador.

La transferencia a través de lenguajes podría ser una técnica para extraer estructuras ontológicas de textos en español en el dominio del conflicto armado, en particular, futuros trabajos podrían abordar estrategias de *meta-learning* considerando como tareas relacionadas la extracción de entidades en español e idiomas ampliamente analizados que posean grandes conjuntos de datos etiquetados. Esta línea de investigación ha demostrado ser útil y mejorar el desempeño de los sistemas, por ejemplo, en [Chowdhury et al., 2020] entrenan un clasificador a partir de textos en inglés, francés e italiano que son codificados a nivel de sentencia con *Multi-lingual BERT* (mBERT) [Devlin et al., 2019], además aplican *manifold mixup* que es una estrategia de visión computacional para aumentar datos, posteriormente el sistema clasifica documentos en español (que nunca antes había visto) y alcanza un *F-measure* de 85.33 %. Estos resultados son producidos por mBERT que "representa algunas características sintácticas que se superponen entre lenguajes" [Chi et al., 2020, p. 6].

Futuros trabajos pueden examinar las técnicas de validación donde predomina la evaluación basada en *gold standard* que requiere corpus etiquetados con información detallada como relaciones taxonómicas y no taxonómicas [Albukhitan et al., 2017, Gupta et al., 2017, Bel-Enguix et al., 2019], limitando la aplicación a dominios que tengan estos recursos o con capacidad de incurrir en los costos de su construcción. Igualmente, los resultados de la evaluación intrínseca mediante *semantic relatedness* y *word similarity* deben analizarse con cuidado pues las referencias (*i.e.* WordSim 353, SimLex-999, MEN-TR-3k, etc) contienen diferentes relaciones semánticas, es decir, existen pares de términos cuya relación puede ser holónimo/merónimo, hiperónimo/hipónimo, relacionados temáti-

[4]Dadas dos entidades e_i y e_j enlazadas por el conjunto R de relaciones, la suposición *distant supervision* considera que todas las oraciones que contienen a e_i y e_j, serán instancias validas de las R relaciones.

camente, etc [Baroni and Lenci, 2011], por ende, es difícil establecer las carencias de un sistema que obtenga un mal desempeño en estas tareas. Sumado a lo anterior, las referencias no tienen terminología de dominios específicos, por lo anterior no permiten valorar estructuras particulares. Una alternativa es diseñar las bases de referencia, por ejemplo en [Jiang et al., 2020] recolectaron 120 conceptos del dominio biomédico y solicitaron a un grupo de expertos valorar la similitud semántica. No obstante, esta postura es costos al ser necesario acceder a conocimiento de peritos.

Capítulo 9

Generación de nuevo conocimiento

El trabajo de la autora fue financiado mediante el programa Jóvenes Investigadores e Innovadores bajo las convocatorias 775 y 812. Así mismo, los labores de la autora estuvieron financiados por la Universidad Tecnológica de Pereira bajo el proyecto "Propuesta de enriquecimiento ontológico a partir de datos textuales para el idioma español en el dominio del conflicto armado colombiano" código E7-20-1.

A continuación, se lista los productos obtenidos en relación a este trabajo.

- Artículo "Generación semi automática de red semántica para describir comunidades víctimas del conflicto armado colombiano" publicado en la revista Entramado en el volumen 16, no. 1 Enero - Junio 2020, doi: 10.18041/1900-3803/entramado.1.6114
- Participación en el evento Cuarta Jornada de Apropiación Social del Conocimiento y la Feria de semilleros UTP -2019 con el trabajo "Generación de redes de información textual que describan factores socio – culturales de comunidades víctimas del conflicto armado como insumo para la formación de políticas públicas y planes de desarrollo".
- Participación en 10th International Conference on Web Intelligence, Mining and Semantics, WIMS 2020 and 5th International Conference on Real-time Intelligent Systems con el trabajo "Semi-automatic extraction and validation of concepts in ontology learning from texts in Spanish"doi: 10.1145/3405962.3405977
- Participación en 2020 IEEE ANDESCON evento que ocurrio en octubre del presente año. La autora participó con el trabajo "Strengthening human-computer interfaces: an automatic construction and evaluation of an annotated corpus"doi: 10.1109/ANDESCON50619.2020.9272046
- Participación en el evento Quinta Jornada de Apropiación Social del Conocimiento y la Feria de semilleros UTP -2020 con el trabajo "Propuesta para la construcción semi - automática de conceptos coherentes en el aprendizaje ontológico mediante textos en español".

Bibliografía

[Aguilar et al., 2016] Aguilar, C., Acosta, O., Sierra, G., Juárez, S., and Infante, T. (2016). Extraction of definitional contexts from biomedical corpora [extracción de contextos definitorios en el área de biomedicina]. *Procesamiento de Lenguaje Natural*, 57:167–170.

[Al-Aswadi et al., 2020] Al-Aswadi, F., Chan, H., and Gan, K. (2020). Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review*, 53(6):3901–3928.

[Albukhitan et al., 2017] Albukhitan, S., Helmy, T., and Alnazer, A. (2017). Arabic ontology learning using deep learning. *Proceedings - 2017 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2017*, pages 1138–1142.

[Alemán et al., 2019] Alemán, Y., Somodevilla, M., and Vilariño, D. (2019). Similarity metrics analysis for principal concepts detection in ontology creation. *Journal of Intelligent and Fuzzy Systems*, 36(5):4753–4764.

[Alfonseca and Manandhar, 2002] Alfonseca, E. and Manandhar, S. (2002). Improving an ontology refinement method with hyponymy patterns. *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 235–239.

[Ali and Melton, 2018] Ali, I. and Melton, A. (2018). Semantic-based text document clustering using cognitive semantic learning and graph theory. *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, pages 243–247.

[Ali et al., 2019] Ali, M., Fathalla, S., Ibrahim, S., Kholief, M., and Hassan, Y. (2019). Cloe: a cross-lingual ontology enrichment using multi-agent architecture. *Enterprise Information Systems*, 13(7-8):1002–1022.

[Alzate, 2010] Alzate, M. L. (2010). Significado de las violencias locales en un mundo globalizado. *Espacio Abierto: Cuaderno Venezolano de Sociología*, 19:505.

[Alzate and Romo, 2014] Alzate, M. L. and Romo, G. (2014). El enfoque de la gobernanza y su recepción en el marco gubernativo actual de las sociedades latinoamericanas. *Opinión Pública*, 20:480 – 495.

[Amoualian et al., 2017] Amoualian, H., Lu, W., Gaussier, E., Balikas, G., Amini, M.-R., and Clausel, M. (2017). Topical coherence in lda-based models through induced segmentation. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1799–1809.

[Asim et al., 2018] Asim, M., Wasim, M., Khan, M., Mahmood, W., and Abbasi, H. (2018). A survey of ontology learning techniques and applications. *Database*, 2018(2018).

[Baroni et al., 2014] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. Association for Computational Linguistics.

[Baroni and Lenci, 2011] Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. GEMS '11, page 1–10. Association for Computational Linguistics.

[Bel-Enguix et al., 2019] Bel-Enguix, G., Gómez-Adorno, H., Reyes-Magaña, J., and Sierra, G. (2019). Wan2vec: Embeddings learned on word association norms. *Semantic Web*, 10(6):991–1006.

[Blair, 2012] Blair, E. (2012). *Un itinerario de investigación sobre la violencia : contribución a una sociología de la ciencia*. Editorial Universidad de Antioquia, 1 edition.

[Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

[Brewster et al., 2004] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation. *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, pages 641–644.

[Buitelaar et al., 2005] Buitelaar, P., Cimiano, P., and Magnini, B. (2005). *Ontology learning from text: methods, evaluation and applications*, volume 123. IOS press.

[Bunk and Krestel, 2018] Bunk, S. and Krestel, R. (2018). Welda: Enhancing topic models by incorporating local word context. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 293–302.

[Carley et al., 2012] Carley, K., Bigrigg, M., and Diallo, B. (2012). Data-to-model: A mixed initiative approach for rapid ethnographic assessment. *Computational and Mathematical Organization Theory*, 18(3):300–327.

[Chang et al., 2009] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, pages 288–296.

[Chen et al., 2012] Chen, S.-Y., Chang, C.-N., Nien, Y.-H., and Ke, H.-R. (2012). Concept extraction and clustering for search result organization and virtual community construction. *Computer Science and Information Systems*, 9(1):323–355.

[Chi et al., 2020] Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[Chowdhury et al., 2020] Chowdhury, J. R., Caragea, C., and Caragea, D. (2020). Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 292–298. Association for Computational Linguistics.

[Christensen and Laegreid, 2005] Christensen, T. and Laegreid, P. (2005). El estado fragmentado: los retos de combinar eficiencia, normas institucionales y democracia. *Gestión y Política Pública*, 14. 3.

[Cimiano, 2006] Cimiano, P. (2006). Ontology learning and population from text: Algorithms, evaluation and applications. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, pages 1–347.

[Cimiano and Völker, 2005] Cimiano, P. and Völker, J. (2005). Text2onto a framework for ontology learning and data-driven change discovery. *Lecture Notes in Computer Science*, 3513:227–238.

[Clark et al., 2012] Clark, M., Kim, Y., Kruschwitz, U., Song, D., Albakour, D., Dignum, S., Beresi, U., Fasli, M., and De Roeck, A. (2012). Automatically structuring domain knowledge from text: An overview of current research. *Information Processing and Management*, 48(3):552–568.

[Dang and Nguyen, 2018] Dang, T. and Nguyen, V. T. (2018). ComModeler: Topic Modeling Using Community Detection. In Tominski, C. and von Landesberger, T., editors, *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association.

[Deerwester et al., 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Degbelo, 2017] Degbelo, A. (2017). A snapshot of ontology evaluation criteria and strategies. *ACM International Conference Proceeding Series*, 2017-September:1–8.

[Dellschaft and Staab, 2008] Dellschaft, K. and Staab, S. (2008). Strategies for the evaluation of ontology learning. *Frontiers in Artificial Intelligence and Applications*, 167(1):253–272.

[Denaux and Gomez-Perez, 2019] Denaux, R. and Gomez-Perez, J. (2019). Vecsigrafo: Corpus-based word-concept embeddings. *Semantic Web*, 10(5):881–908.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

[Dey et al., 2018] Dey, P., Chatterjee, A., and Roy, S. (2018). Knowledge based community detection in online social network. *2018 10th International Conference on Communication Systems and Networks, COMSNETS 2018*, 2018-January:637–642.

[Dugué and Perez, 2015] Dugué, N. and Perez, A. (2015). Directed Louvain : maximizing modularity in directed networks. Research report, Université d'Orléans.

[Edison and Carcel, 2021] Edison, H. and Carcel, H. (2021). Text data analysis using latent dirichlet allocation: an application to fomc transcripts. *Applied Economics Letters*, 28(1):38–42.

[Eissa et al., 2018] Eissa, A., El-Sharkawi, M., and Mokhtar, H. (2018). Towards recommendation using interest-based communities in attributed social networks. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pages 1235–1242.

[Erekhinskaya et al., 2020] Erekhinskaya, T., Morris, M., Strebkov, D., and Moldovan, D. (2020). Leveraging ontologies for natural language processing in enterprise applications. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11878 LNCS:79–85.

[Espinosa, 2018] Espinosa, L. (2018). Tesauro con enfoque diferencial sobre graves violaciones a los DDHH e infracciones al DIH con ocasión del conflicto armado colombiano. Technical report, Centro Nacional de Memoria Historica.

[Farreres et al., 2010] Farreres, J., Gibert, K., Rodríguez, H., and Pluempitiwiriyawej, C. (2010). Inference of lexical ontologies. the leoni methodology. *Artificial Intelligence*, 174(1):1–19.

[Flouris et al., 2008] Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., and Antoniou, G. (2008). Ontology change: Classification and survey. *Knowledge Engineering Review*, 23(2):117–152.

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[Galicia-Haro and Gelbukh, 2014] Galicia-Haro, S. and Gelbukh, A. (2014). Extraction of semantic relations from opinion reviews in spanish. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8856:175–190.

[García, 2014] García, M. (2014). Cultivos ilícitos y confianza institucional en Colombia. *Política y gobierno*, 21:95 – 126.

[Gillani and Ko, 2015] Gillani, S. and Ko, A. (2015). Incremental ontology population and enrichment through semantic-based text mining: An application for it audit domain. *International Journal on Semantic Web and Information Systems*, 11(3):44–66.

[Gruber, 1993] Gruber, T. (1993). *A translation approach to portable ontology specifications*, volume 5.

[Guarino et al., 2009] Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? *Handbook on Ontologies*, pages 1–17.

[Gupta et al., 2017] Gupta, N., Podder, S., Annervaz, K., and Sengupta, S. (2017). Domain ontology induction using word embeddings. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, pages 115–119.

[Gutiérrez-Batista et al., 2018] Gutiérrez-Batista, K., Campaña, J., Vila, M.-A., and Martin-Bautista, M. (2018). An ontology-based framework for automatic topic detection in multilingual environments. *International Journal of Intelligent Systems*, 33(7):1459–1475.

[Gómez-Pérez, 2004] Gómez-Pérez, A. (2004). Ontology evaluation. *Handbook on Ontologies*, pages 251–273.

[Hicks, 2017] Hicks, A. (2017). Metrics and methods for comparative ontology evaluation. *Ciencia da Informacao*, 46(1):34–42.

[Hlomani and Stacey, 2014a] Hlomani, H. and Stacey, D. (2014a). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web and Information Systems*, 1(5):1–11.

[Hlomani and Stacey, 2014b] Hlomani, H. and Stacey, D. (2014b). An extension to the data-driven ontology evaluation. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014*, pages 845–849.

[Horta and Campello, 2015] Horta, D. and Campello, R. (2015). Comparing hard and overlapping clusterings. *Journal of Machine Learning Research*, 16:2949–2997.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:328–339.

[Huang et al., 2020] Huang, W., Mao, Y., Yang, Z., Zhu, L., and Long, J. (2020). Relation classification via knowledge graph enhanced transformer encoder. *Knowledge-Based Systems*, 206.

[Hutchins and Benham-Hutchins, 2010] Hutchins, C. and Benham-Hutchins, M. (2010). Hiding in plain sight: Criminal network analysis. *Computational and Mathematical Organization Theory*, 16(1):89–111.

[Ilgen and Hulin, 2000] Ilgen, D. and Hulin, C. (2000). *Computational Modeling of Behavior in Organizations: The Third Scientific Discipline*. American Psychological Association, 1 edition.

[Jia et al., 2018] Jia, C., Carson, M., Wang, X., and Yu, J. (2018). Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition*, 76:691–703.

[Jiang et al., 2020] Jiang, S., Wu, W., Tomita, N., Ganoe, C., and Hassanpour, S. (2020). Multi-ontology refined embeddings (more): A hybrid multi-ontology and corpus-based semantic representation model for biomedical concepts. *Journal of Biomedical Informatics*.

[Kalyvas, 2001] Kalyvas, S. (2001). La violencia en medio de la guerra civil: esbozo de una teoría. *Análisis Político*, 0(42):3–25.

[Kamoun and Ben Yahia, 2012] Kamoun, K. and Ben Yahia, S. (2012). Automatic approach for ontology evolution based on stability evaluation. *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies*, pages 452–455.

[Khadir et al., 2021] Khadir, A., Guessoum, A., and Aliane, H. (2021). Ontological relation classification using wordnet, word embeddings and deep neural networks. *Lecture Notes in Networks and Systems*, 156:136–148.

[Kingdon, 1984] Kingdon, J. (1984). *Agendas, Alternatives, and Public Policiesia*. Little Brown, 1 edition.

[Knoell et al., 2017] Knoell, D., Atzmueller, M., Rieder, C., and Scherer, K. (2017). A scalable framework for data-driven ontology evaluation. *CEUR Workshop Proceedings*, 1821:97–106.

[Konys, 2019] Konys, A. (2019). Knowledge repository of ontology learning tools from text. *Procedia Computer Science*, 159:1614–1628.

[Korenčić et al., 2018] Korenčić, D., Ristov, S., and Šnajder, J. (2018). Document-based topic coherence measures for news media text. *Expert Systems with Applications*, 114:357–373.

[Lancichinetti and Fortunato, 2012] Lancichinetti, A. and Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, 2.

[Lane et al., 2019] Lane, H., Howard, C., and Hapke, H. M. (2019). *Natural Language Processing in Action. Understanding, analyzing, and generating text with Python*. Manning Publications Co.

[Leicht and Newman, 2008] Leicht, E. and Newman, M. (2008). Community structure in directed networks. *Physical Review Letters*, 100(11).

[Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308. Association for Computational Linguistics.

[Levy et al., 2015] Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

[Li et al., 2013] Li, X., Chen, J., and Zaiane, O. (2013). Text document topical recursive clustering and automatic labeling of a hierarchy of document clusters. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7819 LNAI(PART 2):197–208.

[Li et al., 2020] Li, Z., Nie, J.-Y., Wang, B., Du, P., Zhang, Y., Zou, L., and Li, D. (2020). Meta-learning for neural relation classification with distant supervision. *International Conference on Information and Knowledge Management, Proceedings*, pages 815–824.

[Lin et al., 2016] Lin, P.-L., Huang, P.-W., and Li, C.-Y. (2016). A validity index method for clusters with different degrees of dispersion and overlap. *Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016*, pages 222–229.

[Liu and Alsaadi, 2020] Liu, Y. and Alsaadi, F. (2020). A novel way to build stock market sentiment lexicon. *Communications in Computer and Information Science*, 1179 CCIS:350–361.

[Liu et al., 2013] Liu, Y., Borhan, N., Luo, X., Zhang, H., and He, X. (2013). Association link network based core events discovery on the web. *Proceedings - 16th IEEE International Conference on Computational Science and Engineering, CSE 2013*, pages 553–560.

[Lü et al., 2016] Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., and Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, 650:1–63.

[Lü et al., 2011] Lü, L., Zhang, Y.-C., Yeung, C., and Zhou, T. (2011). Leaders in social networks, the delicious case. *PLoS ONE*, 6(6).

[Meijer et al., 2014] Meijer, K., Frasincar, F., and Hogenboom, F. (2014). A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*, 62:78–93.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119. Curran Associates Inc.

[Mimno et al., 2011] Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 262–272.

[Mishra and Jain, 2015] Mishra, S. and Jain, S. (2015). A study of various approaches and tools on ontology. *Proceedings - 2015 IEEE International Conference on Computational Intelligence and Communication Technology, CICT 2015*, pages 57–61.

[Mosharraf and Taghiyareh, 2017] Mosharraf, M. and Taghiyareh, F. (2017). Domain specific ontology enrichment using public knowledge resources. *2016 8th International Symposium on Telecommunications, IST 2016*, pages 607–611.

[Nkisi-Orji et al., 2019] Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K.-Y., and Heaven, R. (2019). Ontology alignment based on word embedding and random forest classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11051 LNAI:557–572.

[Noy and McGuinness, 2001] Noy, N. and McGuinness, D. (2001). "ontology development 101: A guide to creating your first ontology". Technical report, CStanford Knowledge Systems Laboratory.

[Ochoa et al., 2013] Ochoa, J., Valencia-García, R., Perez-Soltero, A., and Barceló-Valenzuela, M. (2013). A semantic role labelling-based framework for learning ontologies from spanish documents. *Expert Systems with Applications*, 40(6):2058–2068.

[Petasis et al., 2011] Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E. (2011). Ontology population and enrichment: State of the art. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6050:134–166.

[Pfeffer and Carley, 2012] Pfeffer, J. and Carley, K. (2012). Rapid modeling and analyzing networks extracted from pre-structured news articles. *Computational and Mathematical Organization Theory*, 18(3):280–299.

[Ping and Chen, 2018] Ping, Q. and Chen, C. (2018). Litstoryteller+: an interactive system for multi-level scientific paper visual storytelling with a supportive text mining toolbox. *Scientometrics*, 116(3):1887–1944.

[Qiu et al., 2020] Qiu, J., Chai, Y., Tian, Z., Du, X., and Guizani, M. (2020). Automatic concept extraction based on semantic graphs from big data in smart city. *IEEE Transactions on Computational Social Systems*, 7(1):225–233.

[Qiu et al., 2018] Qiu, J., Qi, L., Wang, J., and Zhang, G. (2018). A hybrid-based method for chinese domain lightweight ontology construction. *International Journal of Machine Learning and Cybernetics*, 9(9):1519–1531.

[Reese et al., 2010] Reese, S., Boleda, G., Cuadros, M., Padró, L., and Rigau, G. (2010). Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. pages 1418–1421.

[Reyes-Ortiz, 2019] Reyes-Ortiz, J. (2019). Criminal event ontology population and enrichment using patterns recognition from text. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(11).

[Senel et al., 2018] Senel, L., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T. (2018). Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(10):1769–1779.

[Sfar et al., 2016] Sfar, H., Chaibi, A., Bouzeghoub, A., and Ghezala, H. (2016). Gold standard based evaluation of ontology learning techniques. *Proceedings of the ACM Symposium on Applied Computing*, 04-08-April-2016:339–346.

[Silva and Ribeiro, 2010] Silva, C. and Ribeiro, B. (2010). Inductive Inference for Large Scale Text Classification. *Studies in Computational Intelligence*, 255(January):155.

[Sung et al., 2008] Sung, S., Chung, S., and McLeod, D. (2008). Efficient concept clustering for ontology learning using an event life cycle on the web. *Proceedings of the ACM Symposium on Applied Computing*, pages 2310–2314.

[Syed and Spruit, 2017] Syed, S. and Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, 2018-January:165–174.

[Thenmozhi and Aravindan, 2016] Thenmozhi, D. and Aravindan, C. (2016). An automatic and clause-based approach to learn relations for ontologies. *Computer Journal*, 59(6).

[van Holt et al., 2012] van Holt, T., Johnson, J., Brinkley, J., Carley, K., and Caspersen, J. (2012). Structure of ethnic violence in sudan: A semi-automated network analysis of online news (2003-2010). *Computational and Mathematical Organization Theory*, 18(3):340–355.

[van Holt et al., 2013] van Holt, T., Johnson, J., Carley, K., Brinkley, J., and Diesner, J. (2013). Rapid ethnographic assessment for cultural mapping. *Poetics*, 41(4):366–383.

[Völker et al., 2008] Völker, J., Vrandečić, D., Sure, Y., and Hotho, A. (2008). Aeon - an approach to the automatic evaluation of ontologies. *Applied Ontology*, 3(1-2):41–62.

[Wang et al., 2017] Wang, C., He, X., and Zhou, A. (2017). A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1190–1203.

[Wang et al., 2016] Wang, W., Ning, Y., Rangwala, H., and Ramakrishnan, N. (2016). A multiple instance learning framework for identifying key sentences and detecting events. *International Conference on Information and Knowledge Management, Proceedings*, 24-28-October-2016:509–518.

[Wong et al., 2012] Wong, W., Liu, W., and Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys*, 44(4).

[Xuan et al., 2012] Xuan, J., Jiang, H., Ren, Z., and Zou, W. (2012). Developer prioritization in bug repositories. *Proceedings - International Conference on Software Engineering*, pages 25–35.

[Zablith et al., 2013] Zablith, F., Antoniou, G., D'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D., and Sabou, M. (2013). Ontology evolution: A process-centric survey. *Knowledge Engineering Review*, 30(1):45–75.

[Zafar et al., 2017] Zafar, B., Cochez, M., and Qamar, U. (2017). Using distributional semantics for automatic taxonomy induction. *Proceedings - 14th International Conference on Frontiers of Information Technology, FIT 2016*, pages 348–353.

[Zavitsanos et al., 2010] Zavitsanos, E., Paliouras, G., Vouros, G., and Petridis, S. (2010). Learning subsumption hierarchies of ontology concepts from texts. *Web Intelligence and Agent Systems*, 8(1):37–51.

[Zouaq et al., 2011] Zouaq, A., Gasevic, D., and Hatala, M. (2011). Towards open ontology learning and filtering. *Information Systems*, 36(7):1064–1081.